Jon Øygarden

Norwegian Speech Audiometry

Thesis for the degree of Philosophiae Doctor

Trondheim, April 2009

Norwegian University of Science and Technology
Faculty of Arts
Department of Language and Communication Studies

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Arts
Department of Language and Communication Studies

ISBN 978-82-471-1512-1 (printed ver.)
ISBN 978-82-471-1513-8 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2009:68

Printed by NTNU-trykk

Abstract

A new set of speech audiometry for Norwegian – called “HiST taleaudiometri” – has been developed by the author of this thesis (“HiST” being short for the Norwegian name of Sør-Trøndelag University College and ”taleaudiometri” being Norwegian for speech audiometry). The speech audiometry set consists of five-word sentences, three-word utterances, monosyllabic words, monosyllabic words for testing children and numerals. The process of developing the speech audiometry set is presented in this thesis.

The five-word sentences are of the form Name-verb-numeral-adjective-noun. Hagerman developed this sentence type for Swedish speech audiometry in the 1980s, but for Norwegian the sentences were developed using a new diphone-splitting method. For each word category ten alternatives exist, makings it possible to generate a number of lists with the same phonemic content but with different sentences. A noise was developed from the speech material. This is intended for use together with the speech for the purpose of speech recognition threshold in noise measurements. The material is very suitable for performing repeated measurements on the same person, which is often a requisite for hearing aid evaluation or psychoacoustical testing.

The three-word utterances are of the form numeral-adjective-noun. The words are identical with the last three words used in the five-word sentences. The three-word utterances are intended for speech recognition threshold measurement. The noise developed for five-word sentences can be used together with the three-word utterances for speech recognition threshold in noise measurements.

Monosyllabic word lists were developed mainly for the purpose of measuring maximum speech recognition score or the performance-intensity function. The recorded lists earmarked for testing children were developed by Rikshospitalet University Hospital in Oslo.

The numerals used in the “HiST taleaudiometri” set are the numerals that were recorded by Sverre Quist-Hanssen for his speech audiometry. The numerals are organized in groups of three (digit triplets).

Acknowledgements

Many individuals and institutions have supported the process of developing the Norwegian speech audiometry set “HiST taleaudiometri”: The Ministry of Education and Research granted a scholarship to Sør-Trøndelag University College (HiST) for the purpose of research within audiology. The Norwegian Technological Audiological Society (NTAF-Norsk Teknisk Audiologisk Forening) financed a speech audiometry workshop in 2004 to help set the ambitions for a new Norwegian speech audiometry. The people who sacrificed their time and offered their knowledge during this workshop were Haakon Arnesen, Wim van Dommelen, Laila Glemme, Björn Hagerman, Mari Holmberg, Asbjørn Krokstad, Olav Kvaløy, Einar Laukli, Marte Kristine Lindseth, Olav Mjølsnes, Marit Pedersen, Kolbjørn Slethei, Åshild Spjøtvold, Peter Svensson, Ole Tvete, Arne Vik and Siri Wennberg. Birger Kollmeier and Kirsten Wagener at the University of Oldenburg kindly received me and showed me how they have developed Hagerman’s sentences using new methods. Peder Martin Lysestøl generously read the words and sentences for the recordings of this new material. Lise Utne proofread a preliminary manuscript of this thesis and suggested numerous improvements. The students of BAU2004 and BAU2005 of the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, used preliminary versions of the new speech audiometry set to conduct measurements on a great number of subjects during their practice periods. Furthermore, I am also indebted to the people at the audiological institutions of Norway who enabled the students to perform measurements during their practice periods; the students of BAU2006, who performed measurements on each other as a laboratory study; Martin Sørnes, who conducted measurements of the audio DVD test as part of his bachelor dissertation; as well as the Audiology Programme at HiST and Inger Lise Christensen, who over a period of several years have made possible for me to complete this work. My supervisor Wim van Dommelen at the Department of Language and Communication Studies (ISK), Faculty of Arts, Norwegian University of Science and Technology (NTNU), has supported me and helped me throughout this work.

Thank you to everyone mentioned for your important contributions to my work on this material. The responsibility for any errors found in this work rests solely with me.

Finally, and above all, a special thank you to my dear family Kirsti, Henrik and Marius. Without your support it would have been impossible to complete this work.

Contents

LIST OF FIGURES XV

LIST OF TABLES XXIX

CHAPTER 1 INTRODUCTION 1

1.1 A SHORT HISTORY OF THE DEVELOPMENT OF SPEECH AUDIOMETRY 2
1.2 THE HISTORY OF SPEECH AUDIOMETRY IN NORWAY 6
1.3 REQUIREMENTS FOR NEW SPEECH AUDIOMETRY METHODS 9
1.3.1 WORKSHOP 9
1.4 AIMS OF THESIS 10
1.5 ORGANISATION OF THE THESIS 10

CHAPTER 2 VARIANTS OF HAGERMAN SENTENCES 11

2.1 INTRODUCTION 11
2.1.1 HAGERMAN'S ORIGINAL METHOD 12
2.1.2 IMPROVEMENT OF THE HAGERMAN METHOD: THE WAGENER METHOD 12
2.1.3 IMPROVEMENT OF THE WAGENER METHOD: DIPHONE SPLITTING 13
2.2 METHODS 14
2.2.1 SPEECH MATERIAL 14
2.2.1.1 Selection of names 14
2.2.1.2 Word frequencies 14
2.2.1.3 Phonemic balance 16
2.2.1.4 Word selection 16
2.2.1.5 Recording 17
2.2.2 PREPARATION OF THE STIMULUS MATERIAL 18
2.2.2.1 Naturally read sentences 18
2.2.2.2 Hagerman method 18
2.2.2.3 Wagener method 18
2.2.2.4 Diphone method 20
2.2.2.5 Level adjustments 21
2.2.2.5.1 Uniform with naturally read sentences 21
2.2.2.5.2 Unadjusted 21
2.2.2.5.3 Normalized 22

2.2.2.6 Generating noise 22
2.2.2.7 Preparation of stimulus material for pilot testing 22
2.2.2.8 Preparation of stimulus material for the first field test 23
2.2.2.9 Preparation of stimulus material for the second field test 23
2.2.2.9.1 Level adjustments 23
2.2.2.10 Preparation of the final material 26
2.2.2.10.1 Spectrum of noise and speech material 26
2.2.2.10.2 Adjustments 29
2.2.2.10.3 Sentence material 29
2.2.3 LISTENING TESTS 30
2.2.3.1 Pilot test 30
2.2.3.1.1 The naturalness of material produced by different methods 30
2.2.3.1.2 Speech recognition in noise for material produced by different methods 30
2.2.3.2 First field test: speech recognition in noise 31
2.2.3.2.1 Threshold and slope for words 32
2.2.3.2.2 Familiarization effects 33
2.2.3.2.3 Dialect influence 33
2.2.3.2.4 Noise type 34
2.2.3.3 Second field test: speech recognition without noise 35
2.2.3.3.1 Threshold and slope for words 36
2.2.3.3.2 Verification of the first field test 37
2.2.3.4 First laboratory test: speech recognition threshold measured in hearing level 37
2.2.3.5 Second laboratory test: threshold and slope in noise 38
2.2.3.6 Speech and calibration levels during the listening tests 39
2.2.3.6.1 First field test calibration 39
2.2.3.6.2 Second field test calibration 39
2.2.3.6.3 First laboratory test calibration 40
2.2.3.6.4 Second laboratory test and final speech audiometry material calibration 40
2.3 RESULTS 40
2.3.1 PILOT TEST 40
2.3.1.1 Naturalness 40
2.3.1.2 Speech recognition in noise for material produced by different methods 43
2.3.2 FIRST FIELD TEST 46
2.3.2.1 Word score 46
2.3.2.1.1 Individual words 46
2.3.2.1.2 Word groups 50
2.3.2.2 Training effect 53
2.3.2.3 Dialects 53
2.3.2.4 Noise type 59
2.3.3 SECOND FIELD TEST 60
2.3.3.1 Word score 60
2.3.3.1.1 Individual words 60
2.3.3.1.2 Word groups 63

2.3.3.2 Influence of measurement method 65
2.3.3.3 Verification of the first field test 66
2.3.4 FIRST LABORATORY TEST: THRESHOLD MEASURED IN HEARING LEVEL 68
2.3.5 SECOND LABORATORY TEST: THRESHOLD AND SLOPE IN NOISE 68
2.4 DISCUSSION 69
2.4.1 PILOT LISTENING TESTS 69
2.4.2 FIRST FIELD TEST 70
2.4.3 SECOND FIELD TEST 71
2.4.4 FIRST LABORATORY TEST: THRESHOLD MEASURED IN HEARING LEVEL 71
2.4.5 SECOND LABORATORY TEST: THRESHOLD AND SLOPE IN NOISE 72
2.5 CONCLUSION 72

CHAPTER 3 THREE-WORD UTTERANCES 73

3.1 INTRODUCTION 73
3.1.1 SPONDEE OR NOT SPONDEE 73
3.2 METHODS 75
3.2.1 SPEECH MATERIAL 75
3.2.1.1 Utterances made by using the last three words of the diphone material 75
3.2.1.2 Phonemic balance 76
3.2.2 PREPARATION OF THE STIMULUS MATERIAL 77
3.2.3 LISTENING TESTS 78
3.2.3.1 Second field test: threshold and slope for words without noise, measured in sensation level 78
3.2.3.2 First laboratory test: threshold and slope for words without noise, measured in hearing level 79
3.2.3.3 Second laboratory test: threshold and slope in noise 81
3.2.4 PREPARATION OF THE FINAL MATERIAL 81
3.3 RESULTS 82
3.3.1 SECOND FIELD TEST: THRESHOLD AND SLOPE FOR WORDS WITHOUT NOISE, MEASURED IN SENSATION LEVEL 82
3.3.1.1 Individual words 82
3.3.1.2 Word Groups 84
3.3.2 FIRST LABORATORY TEST: THRESHOLD AND SLOPE FOR WORDS WITHOUT NOISE, MEASURED IN HEARING LEVEL 85
3.3.2.1 Individual words 85
3.3.2.2 Word Groups 87
3.3.2.3 Speech recognition level for calibration 87
3.3.3 SECOND LABORATORY TEST: THRESHOLD AND SLOPE IN NOISE 88
3.4 DISCUSSION 90
3.4.1 SECOND FIELD TEST: THRESHOLD AND SLOPE FOR WORDS WITHOUT NOISE, MEASURED IN SENSATION LEVEL 90
3.4.2 FIRST LABORATORY TEST: THRESHOLD AND SLOPE FOR WORDS WITHOUT NOISE, MEASURED IN HEARING LEVEL 90
3.4.3 SECOND LABORATORY TEST: THRESHOLD AND SLOPE IN NOISE 91
3.5 CONCLUSION 91

CHAPTER 4 MONOSYLLABIC WORDS 93

4.1 INTRODUCTION 93
4.2 METHODS 93
4.2.1 SPEECH MATERIAL 93
4.2.2 SOURCES OF WORDS 95
4.2.2.1 Quist-Hanssen words 95
4.2.2.2 Words for children selected by Rikshospitalet 97
4.2.3 PREPARATION OF THE STIMULUS MATERIAL 98
4.2.4 LISTENING TESTS 99
4.2.4.1 Second field test: detecting easily and less easily recognizable words 99
4.2.4.2 First laboratory test: threshold and slope for words measured in hearing level 99
4.2.4.3 Second laboratory test: masking level 100
4.2.5 PREPARATION OF THE FINAL MATERIAL 101
4.2.5.1 Word inclusions and exclusions 101
4.2.5.2 Phonemic balance 103
4.2.5.3 Mixing strategies 105
4.2.5.4 Level adjustments 108
4.2.5.4.1 Unadjusted 108
4.2.5.4.2 Equivalent level normalized 109
4.2.5.4.3 Loudness normalized 109
4.2.5.4.4 Level adjustments, results 110
4.3 RESULTS 112
4.3.1 SECOND FIELD TEST: DETECTING EASILY AND LESS EASILY RECOGNIZABLE WORDS 112
4.3.2 FIRST LABORATORY TEST: THRESHOLD AND SLOPE FOR WORDS MEASURED IN HEARING LEVEL 112
4.3.3 SECOND LABORATORY TEST: MASKING LEVEL 112
4.4 DISCUSSION 113
4.5 CONCLUSION 113

CHAPTER 5 EVALUATION OF MEASUREMENT STRATEGIES 115

5.1 INTRODUCTION 115
5.1.1 THE PERFORMANCE-INTENSITY FUNCTION 116
5.1.2 BINOMIAL STATISTICS 120
5.2 CONTEXT EFFECTS ON SENTENCE RECOGNITION 122
5.3 METHODS 128
5.3.1 ISO 8253-3 DETERMINATION OF SPEECH RECOGNITION THRESHOLD LEVEL, PROCEDURE A 129
5.3.2 THE HAGERMAN AND KINNEFORS S/N-THRESHOLD METHOD 130
5.3.3 THE HAGERMAN AND KINNEFORS SRT-THRESHOLD METHOD 131
5.3.4 THE BRAND AND KOLLMEIER A1 THRESHOLD METHOD 131
5.3.5 THE BRAND AND KOLLMEIER A2 THRESHOLD AND SLOPE METHOD 133

5.3.6 METHODS WITH A FIXED NUMBER OF ITEMS AT TEST LEVELS WITH CONSTANT INTERVALS 133
5.3.6.1 The constant stimuli method 133
5.3.6.2 The adaptive stimuli method 134
5.3.6.3 The curve-fitting routine 134
5.3.6.4 The counting routine 136
5.4 RESULTS OF SIMULATIONS OF SPEECH AUDIOMETRY MEASUREMENTS 137
5.4.1 ISO 8253-3 DETERMINATION OF SPEECH RECOGNITION THRESHOLD LEVEL, PROCEDURE A 137
5.4.1.1 10 items in test set 137
5.4.1.2 23 items in test set 139
5.4.2 THE HAGERMAN AND KINNEFORS S/N-THRESHOLD METHOD 141
5.4.3 THE HAGERMAN AND KINNEFORS SRT-THRESHOLD METHOD 144
5.4.4 THE BRAND AND KOLLMEIER A1 THRESHOLD METHOD 146
5.4.5 THE BRAND AND KOLLMEIER A2 THRESHOLD AND SLOPE METHOD 148
5.4.6 THE CONSTANT STIMULI METHOD 150
5.4.6.1 3 word sets at 1.5 dB intervals, 141 words/session 150
5.4.6.2 5-word sets at 2.5 dB intervals, 145 words/session 152
5.4.6.3 10 word sets at 7 dB intervals, 110 words/session 154
5.4.6.4 30-word sets at 17 dB intervals, 150 words/session 156
5.4.6.5 50-word sets at 21 dB intervals, 200 words/session 158
5.4.7 SUMMARY OF SIMULATIONS PRESENTED IN "HIST TALEAUDIOMETRI" 160
5.4.7.1 The constant stimuli method with 10 words at 5 dB intervals (150 words/session) 160
5.4.7.2 The 80-20 % adaptive stimuli method with 30 words at 5 dB intervals 161
5.4.7.3 The constant stimuli method with 3 words at 1.5 dB intervals (90 words/session) 162
5.4.7.4 The 80-20 % adaptive stimuli method with 50 words at 10 dB intervals 162
5.4.7.5 The constant stimuli method with 5 words at 2.5 dB intervals (100 words/session) 163
5.4.7.6 The constant stimuli method with 5 words at 2.5 dB intervals (50 words/session) 163
5.4.7.7 The 100-0 % adaptive stimuli method with 9 words at 5 dB intervals 164
5.4.7.8 The constant stimuli method with 3 words at 2 dB intervals (90 words/session) 164
5.4.7.9 The constant stimuli method with 5 words at 3 dB intervals (100 words/session) 165
5.4.7.10 The constant stimuli method with 5 words at 0.75 dB intervals (150 words/session) 165
5.4.8 DETECTION OF DIFFERENCES IN SPEECH RECOGNITION THRESHOLDS 166
5.4.8.1 Introduction 166
5.4.8.2 Detection of differences in SRTs for the simulated speech audiometry procedures 169

5.5 DISCUSSIONS 176
5.5.1 SPEECH RECOGNITION THRESHOLD - SRT 176
5.5.2 SUPRATHRESHOLD MEASUREMENTS WITH MONOSYLLABIC WORDS 177
5.5.3 SIGNAL-TO-NOISE RATIO MEASUREMENTS. 178
5.6 CONCLUSION 181

CHAPTER 6 APPLICATIONS AND RECOMMENDATIONS 183

6.1 "HIST TALEAUDIOMETRI" 183
6.1.1 INTRODUCTION 183
6.1.2 CALIBRATION OF "HIST TALEAUDIOMETRI" 184
6.1.3 THE CONTENTS OF THE "HIST TALEAUDIOMETRI" SET 186
6.1.3.1 CD1 tracks 1-10, monosyllabic words 186
6.1.3.1.1 Deployment 186
6.1.3.2 CD1 tracks 11-13, words for children selected by Rikshospitalet. 187
6.1.3.2.1 Deployment 187
6.1.3.3 CD1 tracks 14-34, three-word utterances for speech recognition threshold measurements 187
6.1.3.3.1 Deployment 187
6.1.3.4 CD1 tracks 35-40, quick-speed test, three-word utterances 187
6.1.3.4.1 Deployment 188
6.1.3.5 CD2 tracks 1-11, five-word sentences for speech recognition threshold measurements 188
6.1.3.5.1 Deployment 188
6.1.3.6 CD2 tracks 12-16, quick-speed test, five-word sentences 189
6.1.3.6.1 Deployment 189
6.1.3.7 CD2 tracks 17-36, binaural tests with earphones 189
6.1.3.7.1 Deployment 190
6.1.3.8 CD2 tracks 37-38, monosyllabic numerals (digit triplets). 192
6.1.3.8.1 Deployment 192
6.1.3.9 CD2 tracks 39-63, signals for calibration 192
6.1.3.10 DVD title 1, three-word utterances for free field audiometry 192
6.1.3.10.1 Deployment 193
6.1.3.11 DVD title 1, five-word sentences for free field audiometry 193
6.1.3.11.1 Deployment 193
6.1.3.12 DVD title 2, 5 five-word sentences in noise and reverberation 194
6.1.3.12.1 Deployment 194
6.1.3.13 DVD title 6, five-word sentences for improved measurement accuracy 196
6.1.3.13.1 Deployment 196
6.1.3.14 DVD title 7, Calibration sounds 1 196
6.1.3.15 DVD title 7, Calibration sounds 2 196
6.2 RECOMMENDATIONS FOR MEASUREMENT OF SPEECH RECOGNITION THRESHOLD 197
6.3 RECOMMENDATIONS FOR MEASUREMENT OF MAXIMUM SPEECH RECOGNITION SCORE 197

6.4 RECOMMENDATIONS FOR MEASUREMENT OF SPEECH RECOGNITION FOR HEARING AID EVALUATION 198
6.5 RECOMMENDATIONS FOR MEASUREMENT OF BINAURAL PERFORMANCE 199
6.6 FURTHER WORK 199
6.7 WHAT COULD HAVE BEEN DONE DIFFERENTLY 201
6.8 CONCLUSION 202

APPENDIX A SCORE PROTOCOL FOR FIRST FIELD TEST 203

APPENDIX B RESULTS FROM FIRST FIELD TEST 205

APPENDIX C SCORE PROTOCOLS FOR SECOND FIELD TEST 207

APPENDIX D SCORE PROTOCOLS FOR SECOND LABORATORY TEST SEPTEMBER 2007 211

APPENDIX E MONOSYLLABIC WORDS 221

APPENDIX F NOMENCLATURE FOR FIVE-WORD AND THREE-WORD LISTS 233

REFERENCES 239

List of Figures

Figure 2.1 Word frequencies and Norwegian Google pages for selected words from the UiB ranking. The filled diamonds show the frequencies of the selected words among the 20 million UiB words. The x's show the number of web pages given by a Google search in Norwegian documents for the same selection of words. The open squares show the number of web pages made obtained with the same method on the same date for the words selected for Hagerman sentences. All the data is shown as a function of the words' ranking among the 20 million UiB words. The left vertical line shows that half of the words selected for Hagerman sentences are among the 3000 most common words in the Bergen material. The right vertical line shows that 48 of the 50 words selected are among the 40 000 most common words in the Bergen material. 15

Figure 2.2 The columns show the distribution of phonemes in the 20 000 most frequent words among the University of Bergen material, corrected for the frequency of each word. The line with diamonds shows the distribution of phonemes in the Norwegian Hagerman material. 16

Figure 2.3 The Matlab tool for generating the building blocks needed in Wagener sentences. 19

Figure 2.4 The Matlab tool for generating the building blocks needed in Diphone sentences. 20

Figure 2.5 Histograms showing the distributions of level adjustments for the different parts of the wave files used to generate diphone sentences. The distance between the horizontal lines corresponds to n= 25. 25

Figure 2.6 The thick line shows the third octave spectrum of 100 sentences containing all the recorded words after the final adjustments. The solid thin lines show the minimum and peak values of the same spectrum. The dotted lines in between show (from bottom to top) the spectrum percentiles 2.5, 25, 50, 75 and 97.5. 26

Figure 2.7 The thick line shows the third octave spectrum of speech noise generated from the sentences. The solid thin lines show the minimum and peak values of the same spectrum. The dotted lines in between show (from bottom to top) the spectrum percentiles 2.5, 25, 50, 75 and 97.5. 27

Figure 2.8 The thick line shows the third octave spectrum of 100 sentences containing all the recorded words after the final adjustments. The thin line shows the third octave spectrum of the noise generated from the sentences. 28

Figure 2.9 Per cent of correctly recognized words for the recognition in noise at -5 dB signal-to-noise ratio for sentences generated by 5 different methods. The group that is significantly different from the others is marked with *. 43

Figure 2.10 Spectrogram, peak levels and waveform of two examples of the sentence “Jonas låner to svarte skåler”. The first example is a naturally read sentence, and this is followed by the same sentence made according to Hagerman’s original method. The vertical lines in the bottom panel are inserted at the word boundaries. 44

Figure 2.11 Third-octave spectra of two 10-sentences lists. Solid line: naturally read sentences. Dotted line: sentences made by means of the original Hagerman method. 46

Figure 2.12 The fitted logistic function for the names as a function of signal-to-noise ratio. 47

Figure 2.13 The fitted logistic function for the verbs as a function of signal-to-noise ratio. 48

Figure 2.14 The fitted logistic function for the numerals as a function of signal-to-noise ratio. 48

Figure 2.15 The fitted logistic function for the adjectives as a function of signal-to-noise ratio. 49

Figure 2.16 The fitted logistic function for the nouns as a function of signal-to-noise ratio. 49

Figure 2.17 The speech recognition threshold for all the words. 50

Figure 2.18 The slope in per cent per dB for all the words. 50

Figure 2.19 The fitted logistic function for the word groups as a function of signal-to-noise ratio. 52

Figure 2.20 The fitted logistic function for all the words as a function of signal-to-noise ratio. 52

Figure 2.21 Score for names in the dialect groups. 54

Figure 2.22 Score for verbs in the dialect groups. 54

Figure 2.23 Score for numerals in the dialect groups. 54

Figure 2.24 Score for adjectives in the dialect groups. 55

Figure 2.25 Score for nouns in the dialect groups. 55

Figure 2.26 The thresholds for logistic functions fitted to each dialect group for the different word groups. 58

Figure 2.27 Slopes s50 for logistic functions fitted to each dialect group for the different word groups. 58

Figure 2.28 The differences between the score for words in test sets A, B, C and D and the mean score for the word as a function of the differences between the SNR for the same words and the mean SNR for the word. The determination coefficient for the fitted line y=-3.16x-0.62 is r2=0.014. 60

Figure 2.29 The fitted logistic function for the names as a function of sensation level. 61

Figure 2.30 The fitted logistic function for the verbs as a function of sensation level. 61

Figure 2.31 The fitted logistic function for the numerals as a function of sensation level. 62

Figure 2.32 The fitted logistic function for the adjectives as a function of sensation level. 62

Figure 2.33 The fitted logistic function for the nouns as a function of sensation level. 63

Figure 2.34 The fitted logistic function for the word groups as a function of sensation level. 64

Figure 2.35 The fitted logistic function for all the words as a function of sensation level. 64

Figure 2.36 Histogram of the difference between the thresholds measured by two methods. 65

Figure 2.37 The slopes of the quick method as a function of the slopes of the full method. 66

Figure 2.38 The left hand columns show the score of sentences with an SNR of -4 dB from the first field test. The right hand columns show the score of the same sentences with an SNR of -5.36 dB in the second field test. 67

Figure 3.1 The cumulative distribution of monosyllabic words (Chapter 4), Quist-Hanssen (Q-H) spondee words, five-word sentences (Chapter 2) and three-word utterances as a function of their ranking among the UiB words. 74

Figure 3.2 Distribution of phonemes: Three-word utterances (solid line with diamonds), the five-word sentences from Chapter 2 (dotted line with triangles) and the 20 000 most frequent words in the University of Bergen material, corrected for the frequency of each word (columns). 77

Figure 3.3 The fitted logistic function for the numerals as a function of sensation level. 83

Figure 3.4 The fitted logistic function for the adjectives as a function of sensation level. 83

Figure 3.5 The fitted logistic function for the nouns as a function of sensation level. 84

Figure 3.6 The fitted logistic function for the word groups and for all the words as a function of sensation level. 85

Figure 3.7 The fitted logistic function for the numerals as a function of hearing level. 86

Figure 3.8 The fitted logistic function for the adjectives as a function of hearing level. 86

Figure 3.9 The fitted logistic function for the nouns as a function of hearing level. 87

Figure 3.10 The fitted logistic function for the word groups and all the words as a function of hearing level. 88

Figure 3.11 Hypothetical subject with slope 16 %/dB simulated measured by fitting a logistic curve to the scores. 28 test items in each set measured at 5 levels. The large panel shows the logistic function for the hypothetical subject, indicated by the thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the fitted logistic curves of the scores. The medium dashed line shows the cumulative distribution of the threshold estimated by the curve fitting routine. The small top left panel shows the histogram of the thresholds obtained during the 500 simulations by the curve fitting routine. The small top middle panel shows the histogram of the estimated slopes. The small top right panel shows the histogram of the estimated rollover parameter. The small bottom panel shows the histogram of the estimated maximum recognition score. The 95 % limits and/or means plus standard deviations of the estimated 89 parameters are indicated.

Figure 4.1 The cumulative distribution of Quist-Hanssen (Q-H) monosyllabic words as a function of their ranking among the UiB words. 96

Figure 4.2 The cumulative distribution of Rikshospitalet's monosyllabic lists for children (RC1 and RC3) and Quist-Hanssen's (Q-H) monosyllabic words as a function of their ranking among the UiB words. 97

Figure 4.3 The cumulative distribution of our new monosyllabic words (NEW), Rikshospitalet's monosyllabic lists for children (RC1 and RC3) and Quist-Hanssen's (Q-H) monosyllabic words as a function of their ranking among the UiB words. 103

Figure 4.4 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The line with diamonds shows the distribution of the new selection of monosyllabic words. 104

Figure 4.5 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The solid line with circles shows the distribution for RC1 monosyllabic words and the dotted line with triangles shows the distribution for RC3 monosyllabic words. 104

Figure 4.6 Score for 10-word groups of monosyllabic words in Appendix C. 107

Figure 4.7 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The solid lines show the distribution of 9 different lists of 50 monosyllabic words. 108

Figure 4.8 Loudness levels estimated for the unadjusted monosyllabic words described in section 4.2.5.4.1. 110

Figure 4.9 Loudness levels estimated for the monosyllabic words normalized to the same equivalent level as described in section 4.2.5.4.2. 111

Figure 4.10 Loudness levels estimated for the monosyllabic words normalized to the same loudness (20 sones) as described in section 4.2.5.4.3. 111

Figure 5.1 Performance-intensity curves for four hypothetical subjects (HS1-HS4). 118

Figure 5.2 The j-factor calculated from listening tests in noise. 123

Figure 5.3 The relationship between recognition probabilities for words and sentences for five-word sentences in the first field test. Horizontal lines above each data point show the number of multiple identical data points. The fitted lines are ps= $(pw)^j$ where j = 4.11 (solid) or j = 1.13 + 4.02·pw (dashed). 124

Figure 5.4 The relationship between recognition probabilities for words and sentences for five-word sentences in the second laboratory test. Horizontal lines above each data point show the number of multiple identical data points. The fitted lines are ps= $(pw)^j$ where j = 4.00 (solid) or j = 0.75 + 3.84·pw (dashed). 124

Figure 5.5 The relationship between recognition probabilities for words and sentences for three-word utterances in the second laboratory test. Horizontal lines above each data point show the number of multiple identical data points. The fitted line is ps= $(pw)^j$ where j = 2.76. 125

Figure 5.6 Four examples of logistic functions generated by the fitting procedure, (lines); and scores, (+) simulated for hypothetical subjects HS1-HS4 (in columns from left to right). 135

Figure 5.7 Simulated measurements of hypothetical subject HS1 according to ISO 8253-3 procedure A with 10 test items at each level. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when “testing” at a specific level. Repeated 138 □

identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 5000 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the number of items tested in each simulation. The 95 % limits for the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown.

Figure 5.8 Results for HS2. Refer to Figure 5.7 for further explanation. 138

Figure 5.9 Results for HS3. Refer to Figure 5.7 for further explanation. 139

Figure 5.10 Results for HS4. Refer to Figure 5.7 for further explanation. 139

Figure 5.11 Simulated measurements of hypothetical subject HS1 according to ISO 8253-3 procedure A with 23 test items at each level. Refer to Figure 5.7 for further explanation. 140

Figure 5.12 Results for HS2. Refer to Figure 5.11 for further explanation. 140

Figure 5.13 Results for HS3. Refer to Figure 5.11 for further explanation. 141

Figure 5.14 Results for HS4. Refer to Figure 5.11 for further explanation. 141

Figure 5.15 Simulated measurements of hypothetical subject HS1 according to the Hagerman and Kinnefors S/N-threshold method. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when “testing” at a specific level. Repeated identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 2500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of number of items tested in each simulation. The 95 % limits for 142 the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown.

Figure 5.16 Results for HS2. Refer to Figure 5.15 for further explanation. 142

Figure 5.17 Results for HS3. Refer to Figure 5.15 for further explanation. 143

Figure 5.18 Results for HS4. Refer to Figure 5.15 for further explanation. 143

Figure 5.19 Simulated measurements of hypothetical subject HS1 according to the Hagerman and Kinnefors SRT-threshold method. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 2500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of number of items tested in each simulation. The 95 % limits for the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown. 144

Figure 5.20 Results for HS2. Refer to Figure 5.19 for further explanation. 144

Figure 5.21 Results for HS3. Refer to Figure 5.19 for further explanation. 145

Figure 5.22 Results for HS4. Refer to Figure 5.19 for further explanation. 145

Figure 5.23 Simulated measurements of hypothetical subject HS1 according to the Brand and Kollmeier A1 threshold method, using 19 five-word sentences. The large panel shows the logistic function for the hypothetical subject as a thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific 146 level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The middle top panel shows the histogram of the thresholds obtained during the 500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the estimated slopes. The 95 % limits for the threshold and the slope plus mean and standard deviation for the threshold and the slope are indicated. Mean and standard deviation for the number of items tested are also shown.

Figure 5.24 Results for HS2. Refer to Figure 5.23 for further explanation. 146

Figure 5.25 Results for HS3. Refer to Figure 5.23 for further explanation. 147

Figure 5.26 Results for HS4. Refer to Figure 5.23 for further explanation. 147

Figure 5.27 Simulated measurements of hypothetical subject HS1 according to the Brand and Kollmeier A2 threshold and slope method, using 20 five-word sentences. The large panel shows the logistic function for the hypothetical subject, indicated with a thick dashed line. Plus signs indicate all the simulated scores obtained when “testing” at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The middle top panel shows the histogram of the thresholds obtained during the 500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the estimated slopes. The 95 % limits for the threshold and the slope plus mean and standard deviation for the threshold and the slope are indicated. Mean and standard deviation for the number of items tested are also shown. 148

Figure 5.28 Results for HS2. Refer to Figure 5.27 for further explanation. 148

Figure 5.29 Results for HS3. Refer to Figure 5.27 for further explanation. 149

Figure 5.30 Results for HS4. Refer to Figure 5.27 for further explanation. 149

Figure 5.31 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 3 test items measured at every level from 5-75 dB in 1.5 dB intervals. The large panel shows the logistic function for the hypothetical subject, indicated by the thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The medium lines show the cumulative distribution of the thresholds estimated by the counting routine, solid line; and the curve-fitting routine, dashed line. The small top left panel shows the histogram of the thresholds obtained during the 500 simulations by the counting routine, solid line; and the curve-fitting routine, bar graph. The small top middle panel shows the histogram of the estimated slopes. The small top right panel shows the histogram of the estimated rollover parameter. The small bottom panel shows the histogram of the estimated maximum recognition score. The 95 % limits and/or means + standard deviations of the estimated parameters are indicated. 150

Figure 5.32 Results for HS2. Refer to Figure 5.31 for further explanation. 151

Figure 5.33 Results for HS3. Refer to Figure 5.31 for further explanation. 151

Figure 5.34 Results for HS4. Refer to Figure 5.31 for further explanation. 152

Figure 5.35 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the 152 counting routine. 5 test items in each set were measured at every level from 5-75 dB in 2.5 dB intervals. Refer to Figure 5.31 for explanation the remaining details.

Figure 5.36 Results for HS2. Refer to Figure 5.35 for further explanation. 153

Figure 5.37 Results for HS3. Refer to Figure 5.35 for further explanation. 153

Figure 5.38 Results for HS4. Refer to Figure 5.35 for further explanation. 154

Figure 5.39 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve fitting routine and the counting routine. 10 test items in each set measured at every level from 5-75 dB in 7 dB intervals. Refer to Figure 5.31 for explanation of the remaining details. 154

Figure 5.40 Results for HS2. Refer to Figure 5.39 for further explanation. 155

Figure 5.41 Results for HS3. Refer to Figure 5.39 for further explanation. 155

Figure 5.42 Results for HS4. Refer to Figure 5.39 for further explanation. 156

Figure 5.43 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 30 test items in each set measured at every level from 1-69 dB in 17 dB intervals. Refer to Figure 5.31 for explanation of the remaining details. 156

Figure 5.44 Results for HS2. Refer to Figure 5.43 for further explanation. 157

Figure 5.45 Results for HS3. Refer to Figure 5.43 for further explanation. 157

Figure 5.46 Results for HS4. Refer to Figure 5.43 for further explanation. 158

Figure 5.47 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 50 test items in each set measured at every level from 14 - 78 dB in 21 dB intervals. Refer to Figure 5.31 for explanation of the remaining details. 158

Figure 5.48 Results for HS2. Refer to Figure 5.47 for further explanation. 159

Figure 5.49 Results for HS3. Refer to Figure 5.47 for further explanation. 159

Figure 5.50 Results for HS4. Refer to Figure 5.47 for further explanation. 160

Figure 5.51 Differences between two simulated measurements on two hypothetical subjects with identical threshold of 35.0 dB. The upper right panel shows the performance-intensity curves for the hypothetical subjects with simulated responses and fitted curves for one of the subjects. The middle top panel shows overlapping histograms of the thresholds for the two subjects. The lower panel shows the cumulative thresholds for the two subjects. The large panel shows a histogram of the differences between the estimated thresholds for the two subjects and the cumulative distribution of the differences estimated by the

Figure F.1 Histogram of differences between five-word sentences with maximum level and minimum level in 10000 lists. 234

Figure F.2 Histogram of differences between three-word utterances with maximum level and minimum level in 100 lists. 236

List of Tables

Table 2.1 The Norwegian words selected for the generation of Hagerman sentences. 17

Table 2.2 Translation of the Norwegian words selected for the generation of Hagerman sentences. 17

Table 2.3 The different types of sentence material used with corresponding label shown in first column. 22

Table 2.4 The required level adjustments for the individual words after the first field test, given in dB. 24

Table 2.5 Example of level adjustments needed for a sentence. 24

Table 2.6 Results in percentage for all the pair comparisons for naturalness. Cells that are positioned symmetrically along the blank diagonal have a total score of 100%. The significant results from this evaluation of naturalness have been given a greyscale background shadow. More pronounced results have darker shadow. 42

Table 2.7 Simplified version of Table 4 where only the differences between level adjustment procedures are shown in per cent. 42

Table 2.8 Simplified version of Table 4 where the details of the level adjustment procedures are removed, so that only the differences between naturally read sentences (ppp), Diphone sentences (dxx), Wagener sentences (wxx) and Hagerman sentences (hxx) are shown in per cent. 43

Table 2.9 The correlation between the dialect groups for the 50 word scores measured with a signal-to-noise ratio of -4 dB. 55

Table 2.10 The results of using Fisher's exact test for all the words 57

Table 2.11 Mean values and standard deviations of the thresholds for the subtests in the five-word sentences binaural test measured in signal-to-noise ratio dB. The results of subtests 1-3, which are measured with at least one ear without masking noise, cannot be presented as dB SNR. However, the mean values of recognized words were 50.0, 49.6 and 49.9 for subtests 1-3 respectively. 68

Table 2.12 Comparison of slope and speech recognition thresholds in noise across four languages. 69

Table 3.1 The Norwegian words selected for generating three-word utterances (English translation). 76

Table 3.2 The wave files needed to generate five-word sentences and three-word utterances. 77

Table 5.1 Parameters used to simulate the four hypothetical subjects. 117

Table 5.2 The standard deviation of a list score with various items per list for different expected true scores. A binomial distribution is assumed. 121

Table 5.3 Rule for adjustment of signal-to-noise ratio after each sentence. 130

Table 5.4 Rule for adjustment of speech level after each sentence. 131

Table 5.5 Computed required differences in the thresholds based on the standard deviation from the simulations in section 5.4.1-5.4.7. The $\sqrt{2}\cdot 1.65\cdot sd$ column is the required difference between two situations for sorting them correctly. The $2\cdot\sqrt{2}\cdot 1.65\cdot sd$ (one-tailed) and the $2\cdot\sqrt{2}\cdot 1.96\cdot sd$ (two-tailed) column is the required difference between two situations for registering that they are different with 95 % confidence. 170

Table 5.6 The best procedures for three-word utterances and five-word sentences. 173

Table 5.7 The best procedures for three-word utterances and five-word sentences modified for use with approximately 150 test items. 175

Table 6.1 Speech recognition thresholds and slopes for the different materials in "HiST taleaudiometri" in silence and with noise. 185

Table A.1 Score protocol for first field test. 204

Table B.1 The score results in per cent for each word from the first field test. 206

Table C.1 Score protocol for second filed test, page 1. 208

Table C.2 Score protocol for second filed test, page 2. 209

Table D.1 Score protocol for first laboratory test, page 1. 212

Table D.2 Score protocol for first laboratory test, page 2. 213

Table D.3 Score protocol for first laboratory test, page 3. 214

Table D.4 Score protocol for first laboratory test, page 4. 215

Table D.5 Score protocol for first laboratory test, page 5. 216

Table D.6 Score protocol for first laboratory test, page 6. 217

Table D.7 Score protocol for first laboratory test, page 7. 218

Table D.8 Score protocol for first laboratory test, page 8. 219

Table D.9 Score protocol for first laboratory test, page 9. 220

Table E.1 The monosyllabic words evaluated for selection. 222

Table E.2 Monosyllabic word lists 1-3. 229

Table E.3 Monosyllabic word lists 4-5. 230

Table E.4 Monosyllabic word lists 7-8. 231

Table F.1 The words used to generate five-word sentences. 233

Table F.2 List number for five-word sentences selected and corresponding LON-number used to generate list. 235

Table F.3 List number for three-word utterances selected and corresponding LON-number used to generate list. 237

Chapter 1

Introduction

The development of a new set of speech audiometry material for Norwegian called “HiST taleaudiometri”[1] is described in this thesis.

Speech audiometry is one of the methods used in audiology to diagnose hearing loss and evaluate various treatments of hearing loss. With this method a list of syllables, words or sentences is presented to a test person at a defined level with or without concurrent noise. The test person responds to what he/she hears and this is recorded and evaluated. The lowest level or signal-to-noise ratio at which the speech signal is intelligible enough to be recognized or identified 50% of the time is the **speech recognition threshold** (**SRT**), also called the **speech reception threshold**. This threshold is traditionally measured with spondaic words. The percentage of words repeated correctly at a given level or signal-to-noise ratio can also be used as a measure of hearing function and is called the **speech recognition score**, **word recognition score** (**WRS**) or the **speech discrimination score**. If the speech recognition score is measured at different levels or signal-to-noise ratios, the **performance-intensity** (**PI**) **function** is measured. If the function is measured with words that are phonetically or phonemically balanced (same proportions of phonemes as in the language, see section 4.2.1) the measured function is often called the **PI-PB function**. The maximum score on the PI-function is called **maximum speech recognition score** (according to ISO 8253-3 (1996), clause 3.12), **maximum speech intelligibility**, **maximum speech discrimination** or **PB$_{\mathbf{max}}$** if it is measured using phonetically balanced material.

Three reasons for performing speech audiometry can be: First, topic diagnosis – to clarify where the location for the hearing damage is. The results of the speech recognition threshold and the maximum speech score have to be evaluated together with the results from other audiological tests. The speech audiometry tests are an important part of the differential diagnostic battery (Thibodeau 2007). Speech audiometry can increase the

[1] HiST is the Norwegian abbreviation for Sør-Trøndelag College, and “taleaudiometri” is Norwegian for speech audiometry

confidence from the results of other tests performed, and indicate further tests needed to be carried out. Second, functional diagnosis – some of the questions that speech audiometry tests can help answering are: How well can this person follow speech and how do noise and reverberation influence the results? Which ear is the best? Does the person benefit from binaural hearing in difficult listening situations? Finally, evaluation of treatments – both the topic diagnosis and the functional diagnosis can reveal need for some sort of rehabilitation, and after performing rehabilitation there is a need for verifying that the goals for the rehabilitation have been accomplished. The treatments can include surgery, fitting of hearing aids or cochlear implants and/or consultation/training etc. Sometimes the evaluation of treatments can be performed with a functional diagnosis, however as part of the process there may exist need for repeated evaluations of different treatments which may require a very high accuracy on the speech audiometry measurements to be able to discern the differences between potential treatments.

The material developed in “HiST taleaudiometri” can have a use for all of these types of diagnosis, and several applications of the material for different purposes have been realized as will be described in Chapter 6.

1.1 A short history of the development of speech audiometry

Speech has been used as an informal test of hearing for a very long time, because conversation becomes difficult both in groups and between two individuals when one of the participants has impaired hearing. Since the early 1800s, more formal testing of hearing using speech signals has developed. Olsen (1990), Bosman (1992), Feldmann (2004) and Wilson and McArdle (2005) have described various aspects related to the history of speech audiometry. Some highlights from this history as described by those scholars are extracted here:

In 1804 Pfingsten (Kiel, Germany) distinguished between three degrees of hearing loss: First, as the most serious loss, hearing loss for vowels. Second, hearing loss for voiced consonants. Finally, hearing loss for unvoiced consonants, which is a milder loss, but also a more common one. Pfingsten used this classification to evaluate a method where galvanic current was applied to the ears of deaf children. In 1801 Grapengiesser in Berlin had reported that this method had been applied with some success.

In Paris, 1821, Itard published Traité des maladies d'oreille et de l'audition, which is the first modern textbook exclusively devoted to diseases of the ear. It describes five classes of increasing hearing loss: First, being able to follow only slow and clear speech. Second, perception of the vowels and some consonants. Third, perception of most of the vowels, but none of the consonants. Fourth, perception of only loud sounds such as thunder. Finally, the fifth category is complete deafness. In 1846 Schmalz (Dresden, Germany) introduced hearing distance as a measure of hearing loss, noting the range within which speech was understood.

Around 1860 Helmholtz demonstrated that vowels are composed of pure tones (Vogel 1993). He was able to show this by synthesizing vowel-like sounds using a tuning-fork apparatus. Helmholtz also constructed an analyzer, which was a tuned set of spherical resonators with two openings where one allowed sounds to enter and the other one was fitted into the ear. With this analyzer he could decompose sung vowels and detected that for each vowel some harmonics were louder in some regions than in other regions of the musical scale. Helmholtz found the same regions of reinforcement for male and female voices, and shifted his emphasis on the basis of this finding. Whereas his previous view had been that vowels were characterized by the relative position of strong harmonics, he now took the position that it was the absolute position of the strong harmonics that characterized each vowel. He also developed a theory of vowels based on the resonance features of the mouth's cavity.

In 1861 Wolf (Frankfurt, Germany) tried to draw up a list of all the speech sounds from low (tongue-R = 16 Hz) to high frequencies (sh = 4096 Hz), and to measure the hearing distance for each sound. Word lists based on these suggestions were produced for some languages. A Gruber quotation in a text-book from 1891 (quoted in Wilson and McArdle 2005, p.80) on the diseases of the ear emphasizes the importance of speech:

> *"Oscar Wolf considers this [speech] the most perfect method of testing the hearing power, inasmuch as it embodies the most delicate shades in the pitch, intensity, and character of sound. Hartmann thinks, on the contrary, that the [speech] test is too complicated to insure accuracy. In any case it [speech measurements] is indispensable, from the fact that nearly every patient seeks relief from disability in respect of it, and therefore for social intercourse. It is desirable, in estimating the degree of perception for speech, to test first of all both ears simultaneously, even though only one be affected; proceeding afterwards to the examination of each [ear] in turn. A separate examination of the hearing power should be made for each ear, even if previous testing by the watch and the tuning-fork has indicated an equally diminished hearing capacity on both sides; since experience shows that the perception for speech is not always deficient in the same measure as that for simple noises and tones. Cases indeed occur in which conversation is best heard on that side on which the watch and tuning-fork are not perceived so*

> *well as on the other, and vice versa. The repetition [repeating] of the test-words gives the best control for the perception of them."*

This shows that at the end of the 19th century speech was considered an important supplement to the frequency-specific information which could be obtained with tuning-forks at that time. Soft and whispered speech was used for diagnostic purposes. In Germany Lichtwitz used Edison's phonograph, which had been invented in 1877, to record speech tests in 1889. This meant that live voice testing could be replaced by a consistent stimulus. Nevertheless, an ideal test stimulus was not achievable because of the poor high-frequency response of the phonograph. In 1904 Bryant (United States of America) recorded monosyllables and the intensity of the speech signal presented through stethoscope tubes was changed during testing by changing the diameter of the tube with a valve. Hearing loss could be expressed in the difference between valve openings for normal-hearing subjects and a hearing-impaired person. The test was never in common use – probably due to the limitations of the phonographic equipment.

In 1910 Campbell and Crandall developed articulation lists consisting of 50 nonsense syllables at the Bell Laboratories. The lists were used to test telephone circuits and each list contained 5 consonant-vowel, 5 vowel-consonant and 40 consonant-vowel-consonant items.

In the 1920s audiometry methods made a major leap forward in the United States of America with the introduction of vacuum tube audiometers as well as recorded test materials. Electronic audiometers had been described in Germany in 1919, and three years later Fowler and Wegel presented the first commercially available audiometer in the United States, the Western Electric 1-A. It was produced in a limited quantity but was used for important studies during the twenties. A smaller and portable version, the 2-A, was later introduced at less than half the price of the very expensive 1-A. Fowler and Wegel also introduced charts, called audiograms, which had the format that is still used today. The audiogram even included an estimate of the speech spectrum. Knudsen (a physicist) and Jones (an otologist) developed an audiometer in Los Angeles in 1924 (Blume and Reeger 1998). This audiometer generated pure tones for air- and bone-conduction testing electronically, included a masking noise source intended for masking the good ear when testing a poor ear and used an attenuator and two vacuum tubes to vary the presented level of speech. The first commercially available speech audiometer was the Western Electric 4-A (1927), introduced by Fletcher from the Bell Laboratories. The audiometer was essentially a phonograph with multiple earphones (Davis and Merzbach 1975). One of the most commonly used tests consisted of digits recorded in groups of three for which the intensity was decreased in 3 dB steps. Fletcher (1929) reports that over a period of three days about 1000 pupils at one of the Public schools in New York were tested with a single phonograph and 40 receivers.

Fletcher and Steinberg's work at the Bell Laboratories built on the earlier efforts of Campbell and Crandall. In a classic article (1929) they describe their methods for designing and implementing tests used in articulation testing. The tests were based on nonsense syllables. Their principles have been followed for over 75 years. The application of these principles at institutions such as the Harvard Psychoacoustic Laboratory and the Deshon General Hospital (Army) during and after World War II evolved into the discipline of Audiology. Nevertheless, nonsense syllables used as stimuli were not considered ideal for clinical speech audiometry, and other stimuli were considered.

In 1947 Hudgins and co-workers at the Harvard Psychoacoustic Laboratory developed two lists of spondees intended for measuring hearing loss for speech. Spondees were selected because they were found to be of more homogenous intelligibility than trochees and iambs. Both syllables in spondees present cues to the listener, and the homogeneity of spondees gives a steep performance-intensity function which assists in measuring hearing thresholds with good accuracy. In 1952 Hirsh and co-workers at the Central Institute for the Deaf revised the spondaic word lists into lists of 36 spondaic words (CID W-1 and W-2) which are still in use.

In 1948 Egan at the Harvard Psychoacoustic Laboratory developed lists of monosyllabic words. Meaningful monosyllabic words are preferable according to Egan, because they represent speech with no syntactic cues – only semantic cues are given to the listener. The lists were phonemically balanced, aiming for the same composition of phonemes in each list as in the English language (PB is short for phonetically balanced, which is the term usually used even though phonemically balanced would be the correct description). The purpose of this balance was to increase the validity of the test for predicting real-life speech perception. However, many of the words included in the Harvard PB lists had a low frequency in English, and Hirsh et al. (1952) revised the lists, which were then published as CID W-22, containing only commonly used words.

This short history, presenting highlights from the development of speech audiometry, was biased towards developments in the United States of America up to the 1950s. These highlighted landmarks have formed the basis for much of the development of speech audiometry material, test equipment and test methods of different kinds that has taken place all over the world since then.

1.2 The history of speech audiometry in Norway

In Oslo around 1950 Sverre Quist-Hanssen developed the most widely used Norwegian speech audiometry material to date. Documentation of this speech audiometry material and the process of developing it are very scarce. The development of speech audiometry materials for Norwegian was intended as the basis for Quist-Hanssen's doctoral dissertation. When the material had been developed and tested the documentation was sent away for statistical treatment. However, all the documentation disappeared, and Quist-Hanssen never received his degree.

Martin Kloster-Jensen counted the phonemes used in Norwegian in 1949 (personal communication, May 8, 1999). This formed the basis for Quist-Hanssen's work. Quist-Hanssen (1965) declares that if speech audiometry measurements need to be related to recognition of everyday speech, then this means that the speech audiometry material must be related to everyday speech. In order for this to be the case, Quist-Hanssen lists the following criteria for the material:

1. The acoustic composition of the word material must be representative of everyday speech.
2. The word material should make low demands on the test person in terms of knowledge of the language and mental effort.
3. Every word in a list must stand out from the rest of the words both acoustically and with respect to meaning.
4. There must be no association between successive words.
5. Words that are easy and difficult to understand must be smoothly distributed in the lists.
6. Different lists with the same types of words must give the same results and deviation.

Quist-Hanssen's (1965) speech audiometry consists of three word types which he described as:

- Digit triplets. Eight monosyllabic numerals in Norwegian were selected: 0 (*null*), 1 (*en*), 2 (*to*), 3 (*tre*), 5 (*fem*), 6 (*seks*), 7 (*sju*) and 12 (*tolv*). The numerals were organized as groups of three. The test was intended for threshold measurements, and should be related to comprehension of easy everyday speech between two persons.
- Spondaic words. These are rare in Norwegian and not as intelligible as the two monosyllabic words which are put together to form the spondee. The selected words had about the same intelligibility and their performance-intensity curve showed a steep slope. The words should be used to measure the speech recognition threshold (SRT).

- Monosyllabic words. Norwegian has many monosyllabic words and more often than not we have to understand all of the speech sounds in order to understand the word. The words are equalized (i.e. the level of each word is adjusted in order to achieve a similar threshold for all of the words), but the spread in threshold will nevertheless be greater than for the spondaic words. The monosyllabic words must be used to measure maximum speech recognition score.

Both the spondaic words and the monosyllabic words were phonetically balanced and equalized. Quist-Hanssen (1970) describes an interesting method used to reduce the spread of the intelligibility threshold of the monosyllabic words, which also applies to subjects with a high-frequency hearing loss. A tape containing equalized words was presented to young normal-hearing subjects through low pass filters with 2500 Hz and 1600 Hz as the upper frequency cut-offs. The increase in level, measured in dB, necessary to make the filtered words intelligible was used as a measure of the loss of intelligibility caused by the two high-tone cut-offs. The words could then be sorted into three main groups: First, words which needed only a moderate increase in the level to become intelligible for both types of high-frequency reduction. Second, words which needed a considerable increase in the level for the 2500 Hz cut-off, but only a slight additional increase in the level for the 1600 Hz cut-off. Finally, words which needed a moderately increased level for the 2500 Hz cut-off, but a very great increase in the level for the 1600 Hz cut-off. In the final lists the words were arranged so that every group of five words was equalized. The words were presented to listeners with different types of hearing capabilities, such as normal hearing, flat hearing loss and different degrees of high-frequency loss.

Quist-Hanssen chose to use the monosyllabic PB words as the standard material for speech recognition threshold and equalized all the word lists (monosyllabic PB words, spondees and digit triplets) in order to achieve a uniform intelligibility threshold (Quist-Hanssen 1966). A justification for this method is that it is easy to judge whether or not the hearing loss is the same for digits, spondees and monosyllabics since the three performance intensity curves will coincide if it is, which can be helpful when diagnosing a hearing loss. However, the method has the drawback of disguising the fact that different speech audiometry materials have different intelligibility.

Quist-Hanssen's speech audiometry is still used extensively throughout most of Norway. It was first released on reel-to-reel tapes, later on compact cassettes and finally on CDs.

In 1975 Kolbjørn Slethei (Bergen, Norway) developed speech audiometry material for the West-Norwegian dialect. Adapting the work of Kloster-Jensen and Quist-Hanssen to requirements based on the West-Norwegian dialect, he developed 12 lists, each with 10 monosyllabic words of the form CV(:)C. He recorded 24 lists, each containing 10 words, where lists 13-24 repeated the words from lists 1-12 but with the words sorted in a different order. Slethei discusses the use of spondaic words for speech recognition threshold measurements, but concludes that:

> *"if a speech audiometry test meets acceptable requirements for predictive validity and face validity, both as a threshold measurement test and as a discrimination test, there are good reasons for using the same test both as a threshold measurement test and a discrimination test. I miss a rationale that explains why bisyllabic words or spondees should be used for threshold measurement and monosyllabic words for the determination of discrimination capability.*
>
> *Based on this that point of view monosyllabic words of the type CV(:)C have been selected as the stimulus material in a test that will function both as a threshold test and as a discrimination test."* (Slethei, 1975, p 25, my translation).

This view is in agreement with the way Quist-Hanssen's speech audiometry has been practiced at some institutions in Norway, where only the monosyllabic words have been used. There are indeed good reasons to be sceptical about the use of spondaic words in Norwegian for threshold measurements since spondees are a relatively rare feature in the Norwegian language. Nevertheless, there are also good reasons to select a material with a steeper performance-intensity function slope than the monosyllabic word material in order to achieve better accuracy when measuring the thresholds. Slethei's speech audiometry is used in the southern part of the west coast of Norway.

The common method for performing speech audiometry in Norway has been to measure points on the performance-intensity function using groups of 10 words at each level. The visually smoothed performance-intensity curve could then be drawn between the measured points.

A few other speech audiometry materials have also been developed in Norway: At Rikshospitalet University Hospital in Oslo a video disc containing IOWA sentences has been developed, and a Hearing In Noise Test (HINT) is in the process of being developed. Universitetssykehuset Nord-Norge HF (The University Hospital of North Norway) in Tromsø is developing a speech audiometry test for the Northern-Sami language.

1.3 Requirements for new speech audiometry methods

The Quist-Hanssen speech audiometry material has many fine qualities, but several people representing different audiological institutions in Norway have suggested that a new speech audiometry is needed. The criticisms directed at the Quist-Hanssen speech audiometry include claims that:

- The selected words are outdated. Many of the words are not in common use today, especially among the spondaic words, but also to some extent among the monosyllabic words.
- It does not include material adapted to measurements on children.
- The sound quality leaves much to be desired. The limitation of the tape recorders used around 1950 in terms of their signal-to-noise ratio is one of the main reasons for this state of affairs.
- The words are pronounced in an old-fashioned way.
- Standardised signal-to-noise measurements cannot be performed because no noise signal has been developed for the material.
- There is insufficient documentation about the development of the material and the levels of the words.
- The intervals between the words do not conform to the requirements of international standards. For some people, the intervals are found to be too short.
- The material is not available in all of the Norwegian dialects.
- The material is not suited to measure the performance of hearing aids.

Not all of these criticisms are fair, since they go beyond the principles selected by Quist-Hanssen as a basis for the development of the material.

1.3.1 Workshop

The Norwegian Technical Audiological Society (NTAF – Norsk Teknisk Audiologisk Forening) sponsored a workshop to help set the ambition for a new Norwegian speech audiometry. Representing different professions in audiology, 18 people from Norway and Sweden participated in the two-day workshop held in February 2004. Rather than consensus, the aim of the workshop was to provide inspiration for further work. Consequently, it involved discussion of many aspects of speech audiometry.

1.4 Aims of thesis

Based on my introductory studies of speech audiometry and the discussions of the workshop I decided to pursue my work in the following order: First, to develop Hagerman's five-word sentences for Norwegian since this type of material can also be used for hearing aid measurements. Second, to select a noise type for use in speech recognition threshold in noise measurements. Third, to evaluate alternatives to spondees as a basis for measuring speech recognition threshold. Next, to select and produce material for maximum speech recognition score measurements. Finally, to evaluate different measurement procedures in order to recommend easy-to-follow test methods (keep it simple!). A prerequisite during the development process was to select material that would also be appropriate for testing children if at all possible. For reasons of time and simplicity, the request to develop speech audiometry materials in different dialects had to be rejected. Rather, the goal was to investigate whether one type of speech audiometry material would function for different dialects.

1.5 Organisation of the thesis

Chapter 2 describes the development of five-word sentences and evaluates Hagerman's original method compared to newer methods. This chapter also describes the development of noise, evaluates results from various dialect regions and presents test results.

In Chapter 3 three-word utterances are presented as an alternative to the spondees and evaluated for speech recognition threshold measurement use.

Chapter 4 presents the selection of monosyllabic words and the process of developing lists to be used for measuring maximum speech recognition score.

In Chapter 5 different measurement procedures are evaluated based on simulations. Some of the results discussed in this chapter are presented in the report (Øygarden 2009) accompanying the speech audiometry disks.

Chapter 6 presents all the tests included on the two CDs and the audio DVD-disk in "HiST taleaudiometri", which is the name given to the end product of this thesis. Recommendations for the deployment of the tests are also presented.

The protocols used for tests performed during the development plus some of the results, a list used in the selection of the monosyllabic words and a nomenclature for the five-word and three-word lists are all included as appendixes.

Chapter 2

Variants of Hagerman sentences

2.1 Introduction

There is need for tests that use natural-sounding speech. The reason is that modern hearing aids can use advanced signal processing strategies. These strategies imply isolating speech from other types of sounds and optimizing the speech for the listener. This signal processing involves long time constants, and may treat isolated words differently from fluent speech. Therefore, we cannot expect to measure the function of the hearing aid correctly when using speech audiometry tests consisting of isolated words. A sentence test may, on the other hand, be well suited for the prediction of perception of natural-sounding speech.

When testing different hearing aids or making adjustments to a hearing aid, we want repeated measurements that help us discriminate between them. We want to repeat many of the measurements in order to evaluate different treatments. The measurements need to be independent of each other and have good reproducibility to allow us to evaluate the results. Making speech audiometry tests which meet these requirements is a complicated matter. We want the speech audiometry lists to be independent of each other, and one way to accomplish this is to use different words in each list. But this may be contradictory to the requirement that the measurements should have good reproducibility. One way of meeting the reproducibility requirement has been to ensure that the lists are phonemically balanced, meaning that the lists have the same distribution of phonemes as that found in the spoken language. This is difficult to achieve for a short list, however, and exhaustive inventories of the distribution of phonemes in the spoken language may not be available. Lyregaard (1997) proposes to use phonemically equalized lists, requiring only that the distribution of phonemes is equal between the different lists. But even this can be a cumbersome task when many lists are needed. Nevertheless, Hagerman found a special way of doing this.

2.1.1 Hagerman's original method

Hagerman (1982) developed a new method for making speech audiometry tests using sentences that minimized the problems outlined in the preceding section. His sentences are constructed so that each list of 10 sentences contains exactly the same 50 words. Hagerman used the same syntactical structure for all of the sentences: (name verb numeral adjective noun). "Ingvild borrows four light plates" is an example of such a sentence.

For each of the elements in a sentence there are 10 alternatives. New sentences are formed by random selection among these alternatives. Hagerman was able to generate a large number of different sentences using his method:

10 names · 10 verbs · 10 numerals · 10 adjectives · 10 nouns = 100 000 sentences

In each list he could take care to combine all the names, verbs, numerals, adjectives and nouns so that each word was used only once. This gives list that are perfectly equalized phonemically. Even on the phonetic level the lists are successfully equalized. Because of the limited amount of words available there is some learning effect when tests are repeated with the different lists. However, the learning effect is mostly initial (Wagener et al. 1999c), and after this initial learning effect has been established a large number of lists can be used to measure the impact of different treatments.

The original Hagerman sentences were based on recordings of the ten sentences where the reader tried to avoid transitions between the words. Each word was then cut out close to the acoustic word boundaries, and new sentences were generated by concatenating the correct words together. These sentences differ from fluent speech in that they have a very staccato rhythm. Wagener and her colleagues therefore developed a new method which made it possible to produce Hagerman sentences without this staccato rhythm.

2.1.2 Improvement of the Hagerman method: the Wagener method

Wagener et al. (1999a) developed a new method enabling them to make more natural-sounding Hagerman sentences. For each of the 10 names Wagener recorded 10 naturally spoken sentences where the name, numeral and noun were kept constant, but with a new verb and a new adjective chosen for each sentence. By using this method Wagener ended up with recordings of 100 natural sentences containing all the possible occurrences of two successive words in the given material. Wagener split each sentence between the word boundaries, except for the last two words in the sentence which could be kept together. This collection of 400 sound files could then be used to generate all the 100 000 possible combinations allowed by the Hagerman sentences. For each word needed in a sentence Wagener could choose from 10 different recordings of the word. She could then select the word from a recording of a sentence where the word was succeeded by the word required by the sentence to be generated. But when she generated the sentence this succeeding word would have to be chosen from a recording of the sentence where this word was succeeded by the next successive word required by the sentence to be generated, and so on.

As an example, suppose she wanted to generate the sentence ***Ingvild borrows four light plates***. She would use *Ingvild* from the recorded sentence ***Ingvild*** *borrows five pretty pens*. She would use *borrows* from the recorded sentence *Malin* ***borrows*** *four pretty gloves*. She would use *four* from the recorded sentence *Malin has* ***four*** *light gloves*. And finally she would use *light plates* from the recorded sentence *Benjamin has seven* ***light plates***.

Using this method Wagener could generate sentences where each word had been read in a sentence where it was succeeded by the correct word. In this way she achieved more natural-sounding sentences than Hagerman, because her sentences contained the correct anticipatory coarticulation between a word and the word succeeding it words. But because each word in the generated sentence is usually taken from a sentence where a different word precedes it, this method will not usually give the correct perseverative coarticulation.

The words were split very close to the beginning of the next word. A weakness of this procedure is that the splitting point can be difficult to choose exactly, and because this point usually belongs to a region where the speech signal is rapidly changing, artefacts may be generated when splicing together words recorded in different sentences. A method for splitting the sentences in a more stable region was required.

2.1.3 Improvement of the Wagener method: diphone splitting

The diphone splitting method uses splitting points more similar to those used in the diphone synthesis method. The same recording could be used as when making Wagener sentences. Instead of splitting between the words, the midpoint of the first vowel in the succeeding word was chosen as the splitting point. This splitting point is located in a very stable phone and can easily be identified. The method gives the correct transition into the succeeding word, which may help us achieve even more natural-sounding sentences than what is possible when using the Wagener method.

2.2 Methods

2.2.1 Speech material

Our selection of words was inspired by the words chosen for the realization of the Swedish and Danish Hagerman sentences. The chosen words had to be familiar to the groups of people who are candidates for this test as we want to measure hearing, not knowledge of words. It is desirable that even small children should know the words.

2.2.1.1 Selection of names

The selection of names was based on the baby naming statistics for Norway presented by Statistics Norway. Five boys' names and five girls' names were selected among the top 100 names for each gender used in Norway during the period 1993-2002. When all the other words had been selected some of the preliminarily chosen names were substituted with others from these groups in order to improve the phonemic balance of the word list.

2.2.1.2 Word frequencies

In order to make this test available to as many people as possible we wanted to use common words. Due to the lack of material describing word frequencies in spoken Norwegian, our evaluation of word frequencies for the selected words was based on words in text. Ranking and frequencies for Norwegian words based on 20 million words of text was available on the web pages of the University of Bergen (2003). The material used in our study was dated 1999 and gave absolute frequencies among 20 million words, and the ranking of these words. The words were selected from books and newspapers and may not represent Norwegian as a spoken language today. Figure 2.1 shows a selection of these words, indicated by filled diamonds. In order to obtain an alternative measurement of word frequencies, we decided to register web page numbers resulting from a Google search in Norwegian documents. The background for this choice was that web pages were expected to represent both a larger basis and more up-to-date material than the texts used in the material provided by the University of Bergen (UiB). The x's in the figure show the numbers for Norwegian web pages found in the Google search for the same words as selected from the Bergen material. If there was good correlation between these materials we would expect the vertical distance between the x's and the diamonds to be constant. The agreement was fairly good except for the rarer words. The open squares show the Hagerman words in terms of the number of web pages in Norwegian documents containing these words according to a Google search as a function of the rank of the words given in the Bergen material.

The vertical line to the left shows that half of the Hagerman words are among the 3000 most common words in the Bergen material. The right vertical line shows that 48 of the 50 words are among the 40 000 most common words according to the Bergen material. The two remaining words are written in the figure. The name *Thea* is among the most commonly used girls' names in Norway lately (see section 2.2.1.1). *Thea* also obtains a higher score on the Google pages than many of the words which have a higher score in the UiB material. The noun *vanter*, which means woollen gloves, has a low score both according to the Google pages and the UiB material. We may assume, however, that this word is a familiar one even among young children in Norway who like to keep their hands warm during the winter.

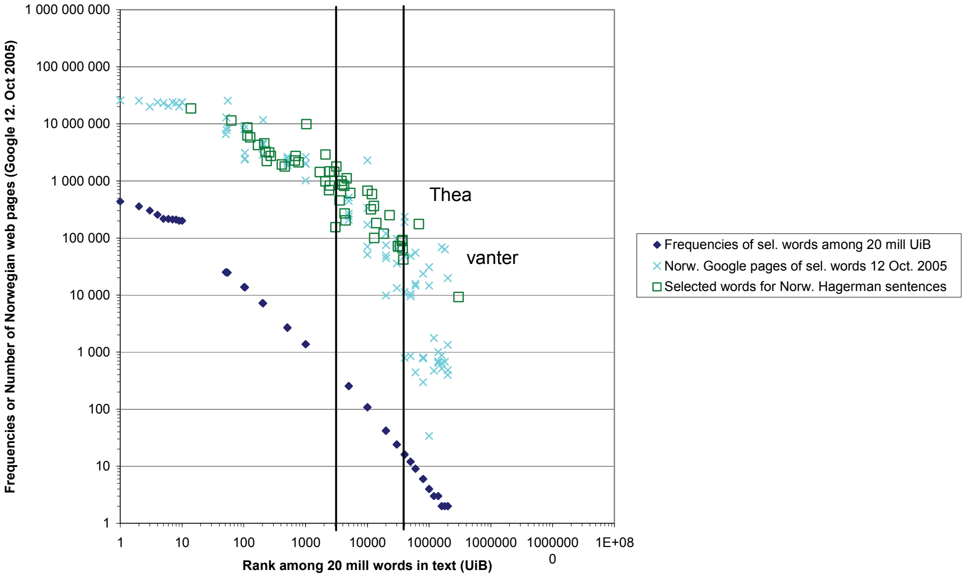


Figure 2.1 Word frequencies and Norwegian Google pages for selected words from the UiB ranking. The filled diamonds show the frequencies of the selected words among the 20 million UiB words. The x's show the number of web pages given by a Google search in Norwegian documents for the same selection of words. The open squares show the number of web pages made obtained with the same method on the same date for the words selected for Hagerman sentences. All the data is shown as a function of the words' ranking among the 20 million UiB words. The left vertical line shows that half of the words selected for Hagerman sentences are among the 3000 most common words in the Bergen material. The right vertical line shows that 48 of the 50 words selected are among the 40 000 most common words in the Bergen material.

2.2.1.3 Phonemic balance

There is no available material describing the phonemic balance of spoken Norwegian. We therefore decided to use the written UiB (Bergen) material as our basis for evaluating phonemic balance. The most commonly used words according to this material could be transcribed with an electronic dictionary developed at the Norwegian University of Science and Technology (NTNU) (Nordgård and Foldvik 2001). These tools were used to make Figure 2.2. The columns show the distribution of phonemes in the Bergen material based on the 20 000 most commonly used words. These data are corrected for the frequency of each word. The line with diamonds shows the distribution of phonemes in the Norwegian Hagerman material.

This produces the best fit between the distributions obtained after reselecting some names, and as the figure shows, the fit is a good one.

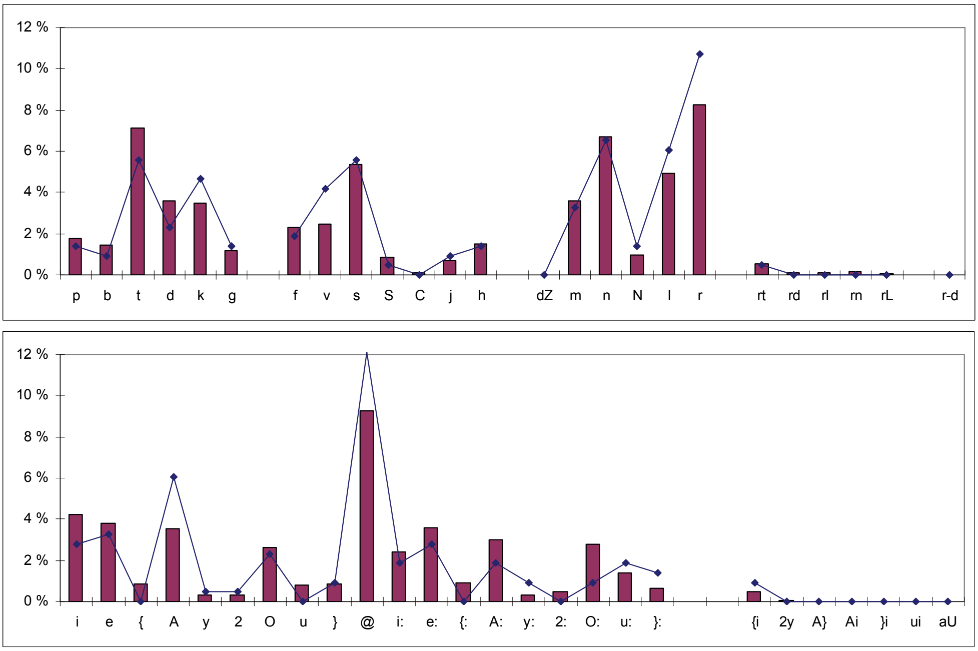


Figure 2.2 The columns show the distribution of phonemes in the 20 000 most frequent words among the University of Bergen material, corrected for the frequency of each word. The line with diamonds shows the distribution of phonemes in the Norwegian Hagerman material.

2.2.1.4 Word selection

Tables 2.1 and 2.2 show the words selected – as described in sections 2.2.1.1 - 2.2.1.3 – for our Norwegian realization of Hagerman sentences and an English translation of these.

Table 2.1 The Norwegian words selected for the generation of Hagerman sentences.

Name	*verb*	*numeral*	*adjective*	*noun*
Hedda	ga	to	gamle	knapper
Ida	grep	tre	hele	boller
Malin	ser	fire	store	vanter
Ingvild	vant	fem	nye	penner
Thea	låner	seks	vakre	kurver
Benjamin	eide	sju	mørke	skåler
Jonas	flytter	åtte	lyse	luer
Thomas	viser	elleve	fine	duker
Magnus	har	tolv	lette	ringer
Eivind	tok	atten	svarte	kasser

Table 2.2 Translation of the Norwegian words selected for the generation of Hagerman sentences.

Name	*verb*	*numeral*	*adjective*	*noun*
Hedda	gave	two	old	buttons
Ida	grabbed	three	whole	muffins
Malin	sees	four	big	gloves
Ingvild	won	five	new	pens
Thea	borrows	six	pretty	baskets
Benjamin	owned	seven	dark	plates
Jonas	moves	eight	bright	caps
Thomas	shows	eleven	fine	tablecloths
Magnus	has	twelve	light	rings
Eivind	took	eighteen	black	boxes

2.2.1.5 Recording

A man of age 62 with an Eastern Norwegian dialect was chosen to read the material. Manuscripts of all the words in Table 2.1 and the 100 sentences as described in section 2.1.2 were prepared with Microsoft Excel. The reader was seated alone in an audiometric room. A microphone Norsonic type 1220, preamplifier Norsonic type 1201 and front end Norsonic type 336 were placed in this room. The recording was made using a DAT- recorder Sony 77ES in the control room. The recording was later transferred digitally to a computer with an SPDIF connection.

2.2.2 Preparation of the stimulus material

The editing of the material was done on PCs using Microsoft Windows XP. Adobe Audition 1.5 and Audacity 1.2 were used for preliminary editing. We developed routines in Matlab 6.1. which we used for further editing and level adjustments. The sound files were saved as 16 bit mono wave files with a sampling frequency of 44 100Hz.

Different methods used for generating sentences are described in sections 2.2.2.1-2.2.2.5. A three-letter label was used to name the different methods. The first letter indicates which one out of four different methods is used for generating the sentence: An initial "p" means naturally read sentence (2.2.2.1); an "h" means Hagerman method (2.2.2.2); a "w" means Wagener method (2.2.2.3); and a "d" means diphone method (2.2.2.4). The last two letters of the three-letter label indicate which method out of four different ones is used for the level adjustment of the sentence: The combination "pp" means a naturally read sentence without adjustment (2.2.2.1); "52" means equivalent to naturally read sentences (2.2.2.5.1); "uj" means without adjustments (2.2.2.5.2); and, finally, "no" means that all elements have been normalized to the same level (2.2.2.5.3).

2.2.2.1 Naturally read sentences

A recording was made of an additional set of natural sentences containing the same words as the Hagerman material. These sentences were meant to serve as a reference for the other sentences to be compared to. During the editing of the material each sentence was saved as an individual file. The only processing done to this material was level adjustment, ensuring that the level was the same for all the sentences. The levels were unweighted equivalent levels calculated directly from the wave files using a Matlab routine. These sentences are labelled ppp.

2.2.2.2 Hagerman method

All the selected words were recorded in isolation. During editing each word was saved as an individual file. To make the sentences the correct files were spliced together.

2.2.2.3 Wagener method

The words were split very close to the beginning of the next word. Our material was produced using a slightly different method than the one applied by Wagener, who gives the following description of her approach: "We attempted to select the point in time for the cutting such that the following word would be perceived as ‘naturally spoken’ if it represented the first word of a new sentence”(Wagener 2003). When generating new sentences the words were shortly ramped (5 ms ramps) and strung together with 5-ms overlap. Instead of using Wagener’s ramping and overlapping method we decided to split the words in positive zero crossings. New sentences could then be generated by simply concatenating the correct files together.

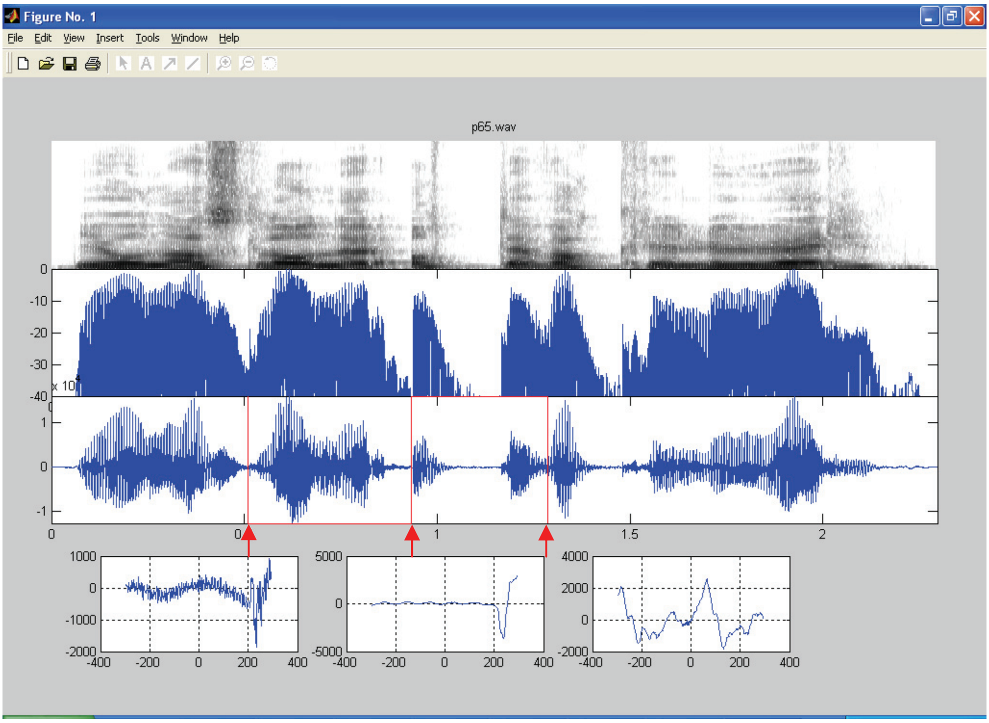


Figure 2.3 The Matlab tool for generating the building blocks needed in Wagener sentences.

The Norwegian sentences were recorded in the same way as Wagener’s material. Each sentence was then saved as an individual file with the sound editor. A specially developed Matlab tool was used for the cutting process. Figure 2.3 is a screen-shot of the sentence *Jonas låner åtte vakre luer* as it looks when using this Matlab tool. From the top we see the spectrogram (0-10 000 Hz), peak levels and waveform aligned to the same time scale. The three vertical red cursors marked with arrows in the waveform window show the splitting points selected for this sentence. The three windows at the bottom show a magnified image of the waveform around the splitting points. Each splitting point was always automatically adjusted to the positive zero crossing close to the cursor. When a cursor was positioned, the tool played the word before and after the cursor, with a 1-second pause in between. This process was repeated until all the words of the sentence could be sounded without any initial or ending artefacts. When we were satisfied with the selections, a right click automatically saved all the parts of the sentence with proper wave file names and fetched the next sentence to be processed.

2.2.2.4 Diphone method

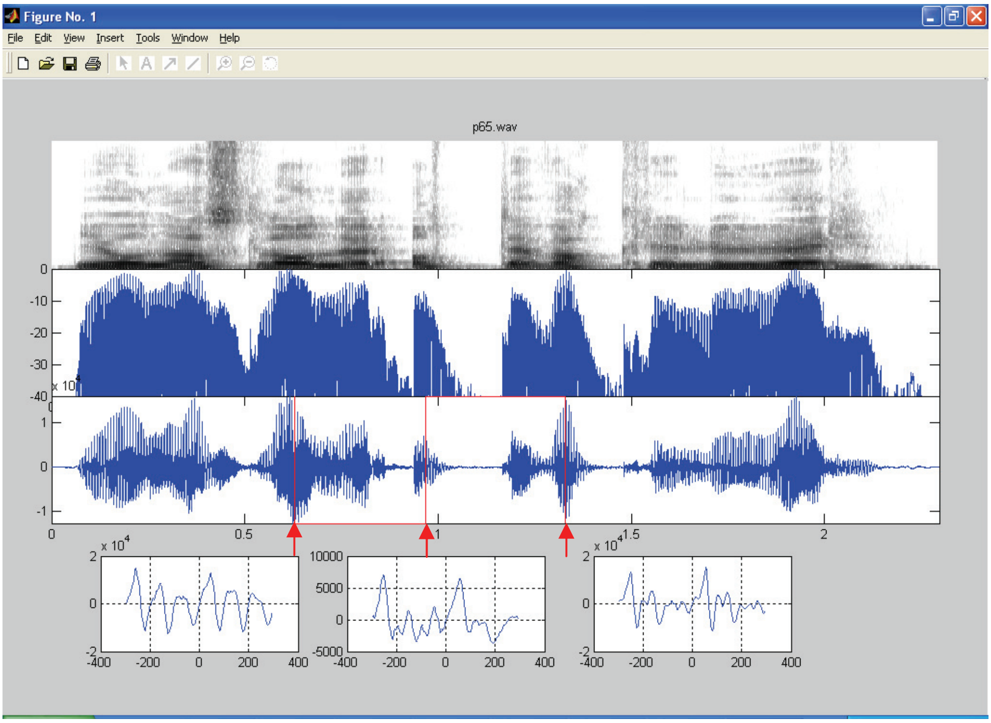


Figure 2.4 The Matlab tool for generating the building blocks needed in Diphone sentences.

With the diphone method the splitting point was selected as the midpoint of the first vowel of the following word. Figure 2.4 is a screen-shot of the sentence used in the preceding paragraph as it looks when fed into a Matlab tool developed to do this splitting. Compared to the preceding figure, different splitting points, marked with arrows, are selected here. The difference between that tool and this one is seen in the three windows at the bottom, where the splitting points are automatically adjusted to the preceding zero crossing of the largest positive amplitude in one period of the vowel. The midpoint was selected visually, evaluated and adjusted if necessary by listening, so that the sound of the vowel could be clearly identified both before and after the splitting point. When we were satisfied with the selections, a right click automatically saved all the parts of the sentence with proper wave file names and fetched the next sentence to be processed.

2.2.2.5 Level adjustments

In the experiments the variants of Hagerman sentences described in sections 2.2.2.2-2.2.2.4 were compared with each other and with the naturally read sentences described in section 2.2.2.1. A problem when making comparisons between these materials is that sentences made according to the procedures in sections 2.2.2.2-2.2.2.4 will not have the same variations in level as a naturally spoken sentence. To minimize this problem we decided to try to reproduce the natural level variation in the more artificially generated sentences. Three different methods were used for level adjustment within the sentences.

2.2.2.5.1 Uniform with naturally read sentences

To minimize any effect of sentence level variation in these tests, we decided to try to make all the sentences similar in terms of level variation. The naturally read sentences were not altered and we attempted to reproduce the level variations found in these sentences in the other sentences. The different methods generate sentences of different lengths even if the words are the same – which makes it very hard to achieve exactly the same level variation in each of the variants of a sentence. As an approximation we created a Matlab routine that divided the sentence in five parts of equal length and adjusted the level of each part so that Hagerman, Wagener and Diphone sentences would match the levels of the naturally read sentences. Abrupt changes in level adjustment were not wanted; therefore the changes in level were varied linearly between the centre points of each of the five equally long parts in a sentence. The level adjustment of the first and the last centre point was kept constant at the start and the end of the sentence. h52, w52 and d52 were all made according to this method.

2.2.2.5.2 Unadjusted

In the test of naturalness the sentences were also tried without any adjustment of the relative levels of the elements in the sentence. ppp, huj, wuj and duj are all made with this method.

2.2.2.5.3 Normalized

Sentences generated with the different methods were also tested with all the elements of a sentence normalized to the same standard level. hno, wno and dno were all made according to this method.

2.2.2.6 Generating noise

Many of the tests for this speech audiometry material require measurements made with background noise. The following procedure was used to produce this noise with the same spectrum as the speech material: Noise was generated by layering a number of sentences from the dno speech material on top of each other. Using Matlab a routine was developed that added together 10 000 sections, each containing a different sentence repeated multiple times without interruptions. The normal level was adjusted to be same as for the sentences and a 30 second segment was saved.

2.2.2.7 Preparation of stimulus material for pilot testing

Using the level adjustments techniques described in section 2.2.2.5 and the different methods for making sentences described in sections 2.2.2.1-2.2.2.4 the ten different types of sentence material presented in Table 2.3 were prepared. Section 2.2.3.1.1 describes how all of these sentence types are used to generate pairs of test sentences. Sentences with a signal-to-noise ratio of -5 dB were also produced from five of the sentence types (to be described in 2.2.3.1.2).

Table 2.3 The different types of sentence material used with corresponding label shown in first column.

ppp	Naturally read sentence without level adjustment.
huj	Hagerman method. Unadjusted.
hno	Hagerman method. Normalized.
h52	Hagerman method. Level adjusted as in a natural sentence.
wuj	Wagener method. Unadjusted.
wno	Wagener method. Normalized.
w52	Wagener method. Level adjusted as in a natural sentence.
duj	Diphone method. Unadjusted.
dno	Diphone method. Normalized
d52	Diphone method. Level adjusted as in a natural sentence.

2.2.2.8 Preparation of stimulus material for the first field test

The results of the pilot listening test supported further development of the material based on the diphone method. More detailed information about the recognition of each word in different situations had to be established. We decided to use the dno sentence material described in the preceding sections, because in this material all the individual elements that were to be concatenated together had been normalized to the same level.

No adjustments were made to the dno material before the first field test. The material was used to generate test sentences with different signal-to-noise ratios (to be described in 2.2.3.2).

2.2.2.9 Preparation of stimulus material for the second field test

The results from the first field test provided information about each word's response when mixed with noise. This knowledge was used to adjust the levels of the diphone words as described in the following section. No other adjustments were made to the material before the second field test.

2.2.2.9.1 Level adjustments

The results from the first field test are presented in section 2.3.2.1.1. The thresholds for the individual words of the fitted logistic function are given. The difference between the threshold of each word and the mean value of the thresholds (-5.36 dB SNR) is the estimated level adjustment of the words that is needed in order to normalize the hearing thresholds. This data is presented in Table 2.4. When performing these level adjustments we have to remember two factors that make this a complex task: First, the elements used to generate the diphone sentences have already been adjusted once for normalization. All these 400 elements used to generate diphone sentences were adjusted to the same level, as reported in section 2.2.2.5.3. In order to find the total amount of adjustment for an element used in building the diphone sentence we need to consider the amount by which each element was already adjusted during the normalization procedure together with the further adjustments given in Table 2.4. The second factor that complicates the level adjustments is that the level adjustments presented in Table 2.4 apply to each individual word. Since we are producing new sentences using the diphone method, each word can be produced by concatenating together two wave files; and each of these wave files contains parts of another word that needs a different amount of level adjustment.

Table 2.5 was made to clarify the requirements of the level adjustment procedure. We have 10 different recordings of each word that need the same amount of adjustment; for 50 words that means 500 adjustments. But the elements used for the diphone synthesis are stored as wave files with the structure shown in row 3 in Table 2.5. A consequence of this is that when adjusting the verb *låner* by 3.0 dB, we need to perform this in two operations. First, we must to increase the level of the last part of wave file 1 by 3.0 dB. Second, we must increase the level of the first part of wave file 2 by the same amount. However, the first part of wave file 1 needed an adjustment of -0.9 dB and the last part of wave file 2 an adjustment of -6.4 dB – these being the requirements for the name *Thea* and the numeral *seks* from Table 2.4.

Table 2.4 The required level adjustments for the individual words after the first field test, given in dB.

Hedda	-0.8	*ga*	2.4	*to*	1.8	*gamle*	-0.5	*knapper*	-2.0
Ida	0.0	*grep*	2.9	*tre*	-0.6	*hele*	2.0	*boller*	-0.9
Malin	-1.3	*ser*	1.1	*fire*	-0.4	*store*	-1.5	*vanter*	-1.8
Ingvild	4.3	*vant*	1.4	*fem*	-0.7	*nye*	-0.6	*penner*	-0.9
Thea	-0.9	*låner*	3.0	*seks*	-6.4	*vakre*	-0.7	*kurver*	2.1
Benjamin	-0.2	*eide*	1.8	*sju*	-2.3	*mørke*	2.4	*skåler*	-2.7
Jonas	-0.1	*flytter*	3.2	*åtte*	1.6	*lyse*	-0.7	*luer*	2.1
Thomas	-2.9	*viser*	-0.4	*elleve*	0.5	*fine*	-0.1	*duker*	0.7
Magnus	-4.0	*har*	-0.4	*tolv*	0.5	*lette*	2.9	*ringer*	-1.0
Eivind	2.0	*tok*	2.1	*atten*	1.1	*svarte*	-3.0	*kasser*	-4.1

Table 2.5 Example of level adjustments needed for a sentence.

Word type	Name	verb		numeral		adjective		noun
Example sentence	*Thea*	*lå-*	*-åner*	*se-*	*-eks*	*va-*	*-akre*	*kurver*
Wave file number	wave file 1		wave file 2		wave file 3		wave file 4	
Adjustment [dB]	-0.9	3.0		-6.4		-0.7		2.1

We have 400 diphone wave files that we use for the diphone synthesis and each of the files needs a different adjustment for its first and last part, meaning that 800 adjustments must be performed. The adjustment procedure requires information about the timing of the border between the first and second part. We already had this timing information for the first four words in the sentences from the production of sentences according to the Wagener method described in section 2.2.2.3, but we had to make new measurements to identify the border between the adjectives and the nouns. All these borders between words were positioned at positive zero crossings, which meant that level adjustments could be performed without transition zones.

The timing and level requirements for the adjustments were computed in an Excel spreadsheet and exported to Matlab. A Matlab routine was developed and used for the adjustments before the second field test.

While inspecting the results from the second field test, we discovered an error in the adjustments. The names, numerals and nouns in the sentences been correctly adjustment, but there was something wrong with the verbs and the adjectives. For instance, the verb *ga* should have an adjustment of 2.4 dB, but the 10 recordings that existed for *ga* had been given the 10 different adjustments for all the verbs shown in column 2 in Table 2.4. This error has been corrected so that the words used in subsequent tests and in the final material have been given the correct level adjustment, but some errors were introduced in the second field test that must be taken into account in the analysis of the results.

The Matlab procedure for adjustments made all the adjustments by multiplying the amplitudes in the selected part of the wave file with a factor computed from Table 2.4. Since all the borders between words and the diphone elements were established at positive zero crossings there was no need for fading.

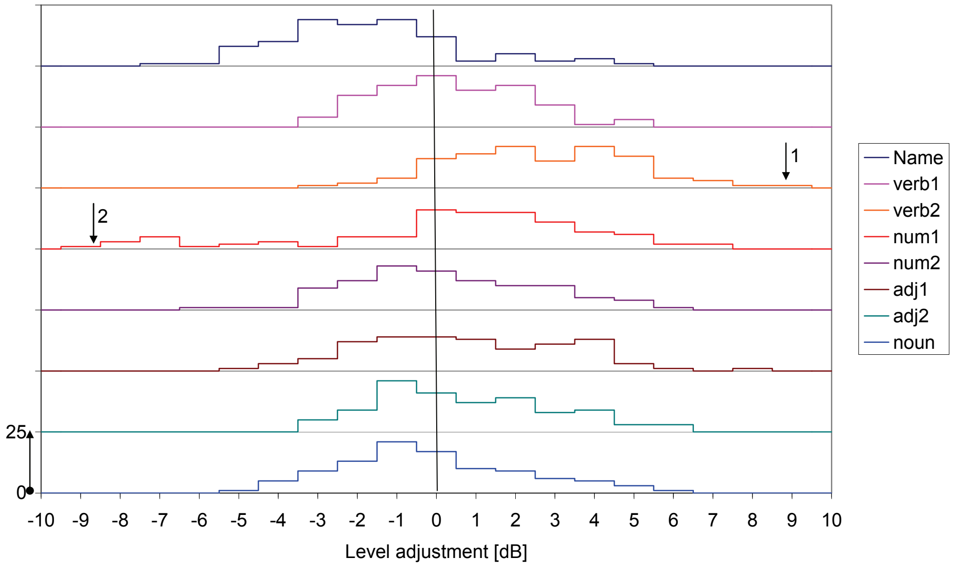


Figure 2.5 Histograms showing the distributions of level adjustments for the different parts of the wave files used to generate diphone sentences. The distance between the horizontal lines corresponds to n= 25.

Figure 2.5 was prepared in order to show the total amount of adjustment made to the elements through the previous normalization of the diphone elements and the adjustment described in this section. The figure shows 8 histograms with 1 dB bins for the 100 adjustments performed on each of the following elements (from top to bottom): Names, verbs part 1, verbs part 2, numerals part 1, numerals part 2, adjectives part 1, adjectives part 2, and finally the nouns.

The largest adjustments are 8.8 dB, indicated in Figure 2.5 by arrow 1 for one instance of the last part of the verb *fly-ytter*; and -8.6 dB, indicated by arrow 2 for one instance of the first part of the numeral *se-eks*. These adjustments were rather large, but 90 % of the adjustments performed were within the range of between -4 dB and +5 dB. Some of the adjustments are noticeable when listening to the material, but this does not seem to cause any substantial reduction in the quality of the material.

2.2.2.10 Preparation of the final material

2.2.2.10.1 Spectrum of noise and speech material

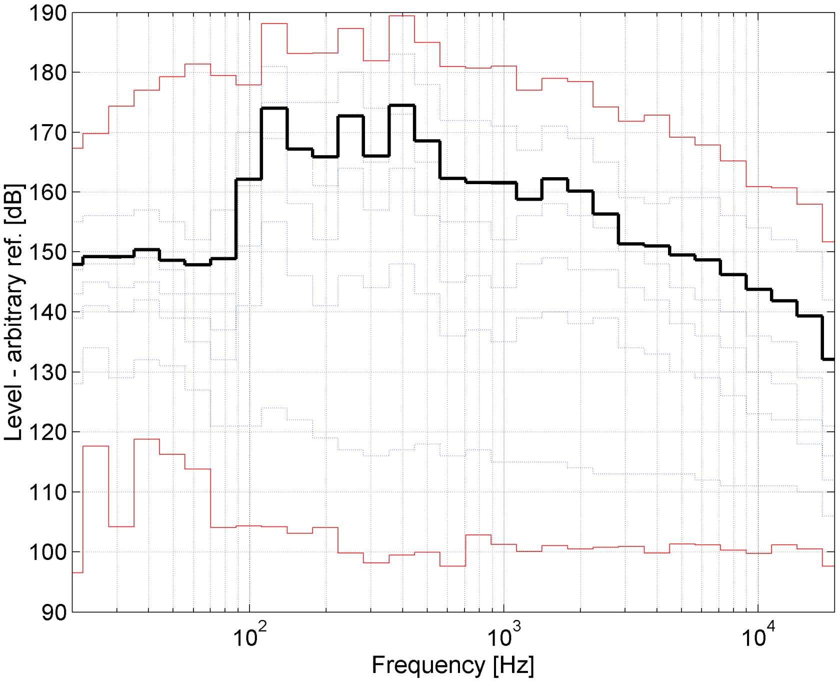


Figure 2.6 The thick line shows the third octave spectrum of 100 sentences containing all the recorded words after the final adjustments. The solid thin lines show the minimum and peak values of the same spectrum. The dotted lines in between show (from bottom to top) the spectrum percentiles 2.5, 25, 50, 75 and 97.5.

The spectra of the sentences and the noise were checked as a quality measure in order to ensure that the processing performed to generate this material had not produced any unanticipated effects. Figures 2.6 and 2.7 show the spectrum of the sentences and the noise respectively. The thick line shows the equivalent spectrum in third octaves. The thin solid lines at the top and bottom show the peak and minimum values in each of the third octaves. The thin dotted lines show the 2.5, 25, 50, 75 and 97.5 percentiles in each third octave. The percentiles are found by grouping the measurements into 1 dB bins. These figures are generated with Matlab code utilizing a 4096 point fast Fourier transform on the wave files with a sampling frequency of 44 100 Hz, which means that the time resolution of the peak-, minimum- and percentile-data is 93ms. This is a sensible value close to the standardized fast time constant of 125 ms for sound level measurements, and the auditory integration times of 100-200 ms found in psychophysical measurements (Moore 2003). The sentence spectra were obtained by analysing a wave file containing the 100 original recorded sentences synthesized with the diphone method after all the adjustments had

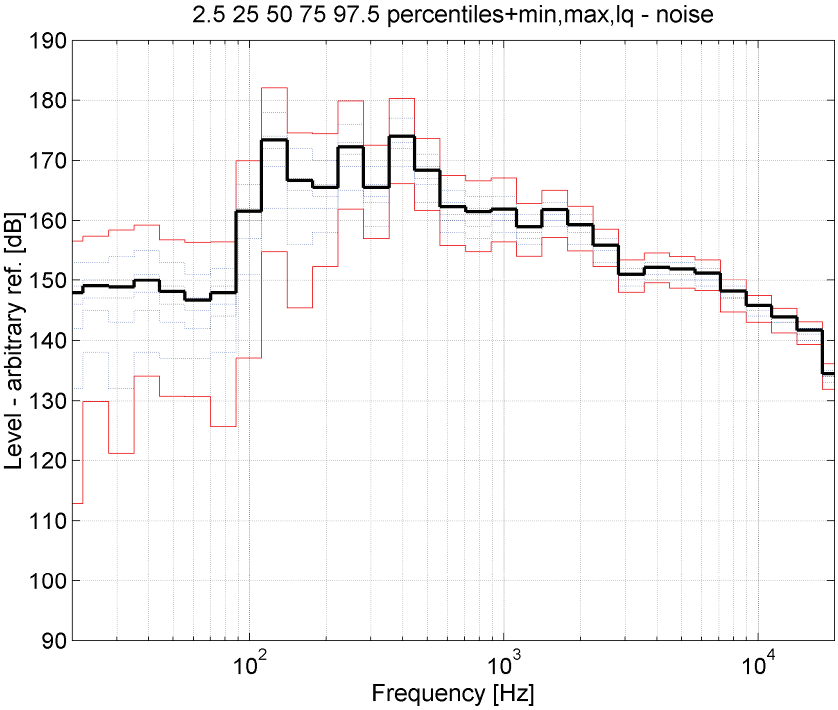


Figure 2.7 The thick line shows the third octave spectrum of speech noise generated from the sentences. The solid thin lines show the minimum and peak values of the same spectrum. The dotted lines in between show (from bottom to top) the spectrum percentiles 2.5, 25, 50, 75 and 97.5.

been made. All the speech sounds necessary to generate the possible 100 000 sentences are contained in this file. The noise spectra were obtained by analyzing the complete 30-second wave file used for generating all noises.

The spectra for the sentences show the wide dynamic range for speech sounds, with a range greater than 60 dB between the 2.5- and 97.5-percentile. The noise has a quite different pattern with a range spanning less than 15 dB if we exclude the lowest frequency bands. This is as expected for a noise signal, but a consequence of this difference in dynamic range is that for part of the time the speech contains spectral energy which is greater than that of the noise.

In Figure 2.8 we compare the equivalent spectra of the speech and the noise. There are small differences between the spectra of the sentences and the noise. Ideally these should be identical. The reason for the discrepancy

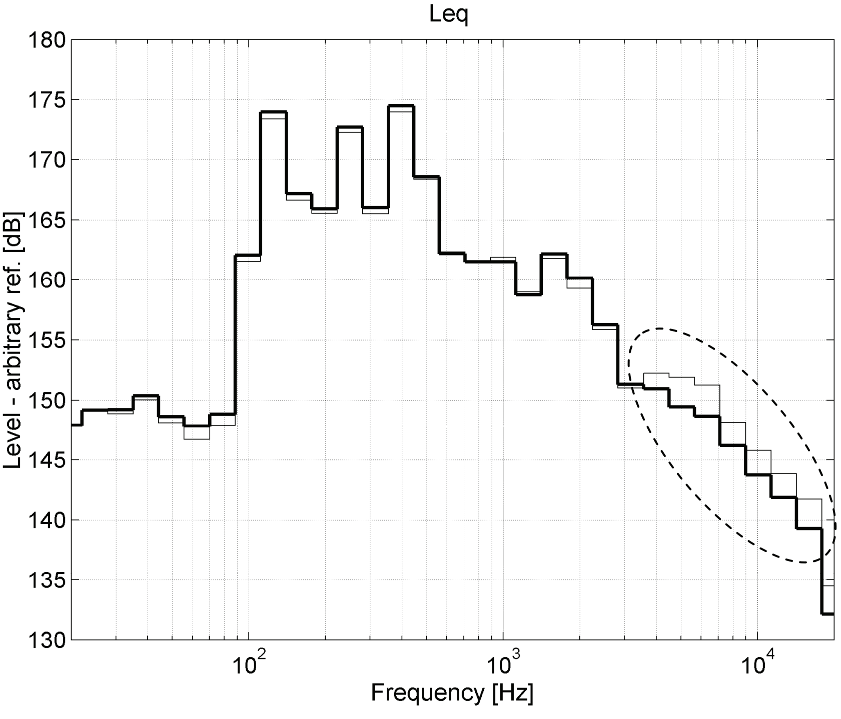


Figure 2.8 The thick line shows the third octave spectrum of 100 sentences containing all the recorded words after the final adjustments. The thin line shows the third octave spectrum of the noise generated from the sentences.

is that the spectrum of the sentence material is obtained after the final adjustments of the word levels; however, the noise was generated from the diphone sentences after the initial normalization but before the final adjustment. As can be seen from Figure 2.8, the differences between the spectra are very small for frequencies below 3500Hz. For the higher frequency bands, indicated by the ellipse drawn with a dashed line in Figure 2.8, the differences are greater, as the sentence spectrum is about 2dB lower in this region. An explanation can be found by inspecting Table 2.4, where we find that the levels were reduced for most of the words containing high frequency s-sounds during the final adjustment.

2.2.2.10.2 Adjustments

As described in section 2.2.2.9.1 an error was revealed in the level adjustments of the verbs and adjectives used in the second field test after the test had been conducted. Correct adjustments in accordance with the intentions described in section 2.2.2.9.1 have since been performed; hence all the words used in the subsequent tests and the final material are adjusted to correct levels.

2.2.2.10.3 Sentence material

400 wave files are available for generating five-word sentences; realization of the 10 000 lists containing all the potential 100 000 sentences which can be made using this method is therefore possible. A nomenclature was developed in order to control the lists and the sentences realized so that repetition of sentences could be avoided. The nomenclature is presented in Appendix F.

Matlab routines were developed for production of new wave files, which typically contained a complete list of ten sentences in one channel and optional noise sequences in the other channel. A list was selected from the nomenclature and the order of the sentences was randomized in a spreadsheet where a text table of the sentences and codes for the sentences were generated. The text table was used in the documentation of the tests. The codes for the sentences were imported into the Matlab routine. The Matlab routine then automatically selected and concatenated the wave files necessary for generating each sentence, adjusted the levels for each sentence according to the test procedure, inserted pauses between the sentences and optionally inserted noise sequences in the other channel in the right places. The noise was fetched from a random start position of the noise wave file described in section 2.2.2.6.

2.2.3 Listening tests

2.2.3.1 Pilot test

2.2.3.1.1 The naturalness of material produced by different methods

A paired comparison was made for some of the sentences. In order to be able to make a pair comparison of the 10 different methods of generating sentences, 45 sentence pairs were needed (10·9/(1·2). This set of 45 sentence pairs was prepared with some sentences used for more than one pair so that these 45 pairs contained 24 different sentences. The pairs were made with a 1-second pause between the two sentences. Each pair consisted of two sentences which contained the same words, but where each was made according to a different choice among the 10 different methods for generating sentences. This made it possible to test each method for making sentences against the 9 other methods.

The test was self-administered on a computer; the test persons were listening binaurally with Sennheiser HD535 headphones connected to the sound card. The level was adjusted to a comfortable listening level. 14 normal-hearing test persons took this test and were asked to evaluate which sentence sounded the more natural in each pair.

All the test persons did this test first so that they were somewhat familiar with the test material before the test of speech recognition in noise.

The binomial distribution was used to evaluate whether or not there were significant differences between the methods. 10 out of the 14 participants had to judge a sentence more natural in a pair to get a significance better than 0.05.

2.2.3.1.2 Speech recognition in noise for material produced by different methods

The 30-second noise wave file developed as described in section 2.2.2.6 was used to make material with a signal-to-noise ratio of -5 dB.

When sentences with a specific signal-to-noise ratio were to be generated, part of the noise was selected from a random starting point within this file and the duration of the sample was 1.5 seconds longer than the sentence to be generated. Before further processing, fading was applied to this clip: 50 ms in and out.. The noise level was kept constant and the level of the sentence was adjusted according to the desired signal-to-noise ratio. The sentence and the noise were then added together, with the noise starting 1 second before the start of the sentence and lasting for 0.5 seconds after its completion, and saved as a file. CDs were made containing the test material.

The test persons were seated in an audiometric booth and the operator on the outside. The test sentences were administered to one ear (the ear chosen by the test person) with a CD player connected to a GN Otometrics Aurical Plus audiometer. TDH 39 earphones were used. The level was adjusted to a comfortable listening level.

Sentences of the 5 variants ppp, h52, w52, d52 and dno type were evaluated at a signal-to-noise ratio of -5dB. The test persons were divided into 5 groups such that the same sentence could be tested for the 5 variants without being used more than once for each test person. 15 test persons participated and listened to 50 sentences each. This gives 750 word tests for each method (15 test persons · 10 test sentences for each method · 5 words in the sentence). The number of correct words was recorded for each test.

ANOVA testing was used to check whether there were significant differences between the groups. If differences were found to exist, post hoc testing would be conducted by using the Student-Newman-Keuls method to evaluate these differences.

2.2.3.2 First field test: speech recognition in noise

Because we needed information about the performance of the dno sentences for further adjustment and fine tuning, a field test was prepared. Students of the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, performed this test during their practice period in January-February 2006.

The first field test had four main objectives: First, it was necessary to obtain knowledge of which of the words were easier or harder to recognize. Second, we wanted to examine the effect of familiarization with the words and test environment, to see whether the scores would improve in the latter part of the test session compared to the scores achieved at the start of the session? Third, would the dialect background of the test persons have any influence on the score? Fourth and finally, we wanted to study effects of the noise, because the test material was produced on compact discs where the noise was temporally fixed to the speech stimuli.

The test material was distributed between four sets, A, B, C and D, each of which contained a CD and a measurement protocol. The score protocol for set A is given in Appendix A. The other protocols have the same structure; only the sentences and signal-to-noise ratios were changed. The sentences with noise were generated by the same procedure as described in section 2.2.3.1.2. The measurement procedure consisted of five different parts. First, two sentences were presented; one without noise and the other with a signal-to-noise ratio of 1 dB. These sentences were repeated if necessary and were used to adjust the sound level to one that was perceived as comfortable. After this acclimatization the real measurements started with a group of ten sentences (numbers 3-12 in the protocol) which had a signal-to-noise ratio of -4 dB. These sentences were the same for all four sets A-D, but the noise was generated independently for each set. Sentences 13-32 were then presented without noise for further acclimatization to the words used in the test. These sentences were also included in all of the sets, A-D. Next, 40 sentences (numbers 33-72 in the protocol) were presented, with a varying signal-to-noise ratio. The sentences were presented randomly within the group, and four out of the 16 signal-to-noise ratios of -12, -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2 and 3 dB were selected for each test set. This ensured that each of the signal-to-noise ratios was used for ten sentences over all test sets combined. Finally, the last group consisted of sentences 73-82 – these were sentences 3-12 repeated, but in a different order.

A total of 107 normal-hearing test persons participated in these measurements, distributed as 30 persons on test set A, 18 on B, 33 on C and, finally, 26 on D. The test ear was chosen by the test person.

2.2.3.2.1 Threshold and slope for words

From the first field test we obtained scoring information for every word at all the signal-to-noise ratios in 1 dB steps between -12 dB and 3 dB. Kollmeier et al. (2008) have showed that the sigmoid function gives a good description of the performance-intensity score for speech audiometry data. The sigmoid function is a special case of the logistic function and can be written as:

$$SI(SNR) = \frac{1}{1+e^{-4s50(SNR-SRT)}} \qquad (2.1)$$

SI is the predicted score for a given signal-to-noise ratio SNR. The parameter s50 represents the slope of the function at the speech recognition threshold SRT (the SNR giving 50% score).

According to Kollmeier's probabilistic model, presented in Wagener (1999b), the intelligibility function of a sentence test depends both on the slopes of the words and the distribution of the SRT values. The intelligibility function can be calculated by a convolution of these factors. In order to achieve a steep $s50_{\text{sentence}}$ slope for the sentences the standard deviation of the words' SRT values must be small, and the mean $s50_{\text{word}}$ slope for the words must be steep. The estimated $s50_{\text{sentence}}$ slope can be calculated by:

$$s50_{sentence} \approx \frac{s50_{word}}{\sqrt{1+\frac{16s50^2_{word}\sigma^2_{SRT}}{\left(\ln\left(2e^{\frac{1}{2}}-1+2e^{\frac{1}{4}}\right)\right)^2}}} \tag{2.2}$$

where σ_{SRT} is the standard deviation of the speech recognition thresholds for the words.

The results of the listening tests were collected in Excel. Subsequently the best fit of a logistic function was obtained for each word with the method of least squares by using the solver which is a standard add-in of the Excel spreadsheet.

2.2.3.2.2 Familiarization effects

When conducting repeated measurements with these lists, we need to be aware of the fact that the test person acquires knowledge about the sentence structure, the words used in the test, the structure of the noise and the test situation. This knowledge grows continuously as the test proceeds and will have an impact on the results we obtain. Hagerman (1984) found that the learning effect associated with these sentence tests will usually not exceed 1dB. Wagener (2003) found a training effect of 2.2 dB difference between the first and last list for test persons listening to 8 test lists, each containing 20 sentences, during a training phase. However, the greatest difference was approximately 1 dB between the first and second list.

In the first field test the test persons will respond to a test list of 10 sentences with a signal-to-noise ratio of -4 dB at the start of the session just after the initial two sentences used to set sound levels. The same sentences will be repeated in a different order at the end of the session. In between these exposures the test persons will have gone through a session which can be categorized as a training phase consisting of, first, 20 sentences without noise followed by 40 sentences with different signal-to-noise ratios.

2.2.3.2.3 Dialect influence

According to Vikør (2001), *"Dialects have a much higher prestige in Norway than in the other Scandinavian countries, and they are used relatively freely in most contexts".* The quotation is found on the web pages of The Norwegian Language Council (Norsk språkråd 2007) and suggests that many Norwegians are very strongly connected with their dialects. From the same source we learn that Norwegian dialects are usually divided into five main groups:

1. Western
2. Northern
3. Eastern
4. Central (i.e. comprising the mountain valleys of the interior)
5. Trønder (of Trøndelag)

Some people have been interested in developing speech audiometry test material for each dialect group. Being able to offer such material would probably come with the satisfaction of pleasing the candidates for speech audiometry measurements. However, generating such diverse materials is far beyond the scope of this work. It is our hope, though, that producing a Norwegian speech audiometry set in only one dialect, while seeking knowledge about how it functions for the other dialect groups, represents a satisfactory solution at this stage. It offers opportunities to use the material for instance when performing national clinical investigations to compare hearing aids at different centres in Norway. Stensby et al. (2002) compared speakers from the Western, Northern, Eastern and Trønder dialect groups in terms of their understandability among both normal-hearing and hearing impaired listeners from the same groups in a similar test situation using sentence tests. A surprising result was that speakers with an Eastern Norwegian dialect background produced the lowest error rates among the listeners irrespective of their dialect group. We used a man with an Eastern Norwegian dialect background for our recordings, as described in section 2.2.1.5.

The dialect of the listener was to be registered during the field test performed by the students during their practice period, which involves working at hearing centres all around Norway. This was an opportunity to collect results from all of the different dialect groups, and we could compare sentences 73-82 which had a signal-to-noise ratio of -4 dB and were included in all of the test sets (A, B, C and D). We could also use the measurements for all the different signal-to-noise ratios on sentences 33-62, and compute logistic functions for the performance of each dialect group in relation to different groups of words.

A total of 107 normal-hearing subjects participated in these measurements, distributed as 39 persons with Western, 23 persons with Northern, 12 persons with Eastern, only 2 with Central and finally 31 persons with Trønder dialect background.

2.2.3.2.4 Noise type

Compact disks are very suitable distribution media for sound needed in speech audiometry tests. All clinical audiometry stations usually have a CD-player available. For the purpose of conducting speech audiometry measurements in noise the speech signal can be recorded on one channel and noise on the other channel. This opens up the opportunity of using the specially developed noise with the same spectrum as the speech when performing speech audiometry. One limitation associated with this method is found in the fact that the speech and noise signals get temporally fixed to each other. A consequence of this is that one specific word in a sentence may get a better signal-to-noise ratio than the rest, and will thus always have a better chance of being recognized than other words. This situation would not have occurred if we had opted for an independent noise generator, which would have allowed the signal-to-noise ratio to vary a little from test to test.

In the first field test we had 4 sets, A-D, containing the same sentences 73-82 with the same signal-to-noise ratios of -4 dB. Each of the sets was produced with a new mixture of speech and noise. The signal-to-noise ratios for all the words in these sentences were measured with Matlab routines and transferred to Excel. Linear regression will be performed to check whether the variations in signal-to-noise ratios have influenced the score.

2.2.3.3 Second field test: speech recognition without noise

A second field test was performed by students of the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, during their practice period in May-June 2006. This field test incorporated test items from the five-word sentences, described in this chapter, plus test items from the three-word utterances described in Chapter 3 and test items from the monosyllabic words described in Chapter 4.

The main objective of using the five-word sentences in the second field test was to acquire information about how the hearing threshold of the individual words varied when measured without background noise. We had adjusted the levels, so we expected rather equal performance for all the words when tested with noise, but it may also be relevant to perform the test around the hearing threshold without noise, and knowledge was needed about any possible differences between the words.

Correct calibration of the levels represents a challenge when performing a test like this one. The students brought their test material to different hearing centres in Norway, and found that some of these lacked the equipment and expertise needed to perform a calibration. Even if the compact disks were set up with the same calibration levels as the ones used for the Quist-Hanssen speech audiometry, small deviations from the correct calibration might exist, potentially invalidating our data. The decision was therefore made not to try to obtain absolute levels for the hearing threshold, but to normalize for each listener’s hearing threshold in relation to these sentences. A sensation level (SL) scale could then be established. This enabled us to evaluate the performance of the individual words according to this sensation level scale, and the test data could then be pooled together without concerns about different calibration levels at the hearing centres.

An example of the scoring protocol can be found in Appendix C. Seven test sets (A, B, C, D, E, F and G) were made. These were uniform in structure but contained different sentences and words. The measurement procedure consisted of speech audiometry divided into four different parts, after initial pure tone audiometric measurements on the ear selected and adjustment of the calibration level on the speech audiometer. The PTA (Pure Tone Average of 500, 1000, 2000 and 4000 Hz) of these initial measurements was used to calculate a start level for the succeeding measurements. The first part of the speech audiometry contained 18 lists of 10 five-word sentences (tracks 2-19 on the CD) without noise. These were used for measuring the performance-intensity function, following the protocol described in the next section. The second part consisted of 15 lists of 10 three-word utterances (tracks 20-34), for measuring the performance intensity function according to the method to be described in section 3.2.3.1. The third part of the procedure was a list of 10 five-word sentences (track 35) with background noise at a signal-to-noise ratio of -5.36 dB. Section 2.2.3.3.2 describes the method used to evaluate these measurements. Finally, the fourth part involved 88-90 monosyllabic words (track 36). The method used for this test will be described in section 4.2.4.1.

33 normal-hearing test persons participated in the second field test, distributed as 7 persons on test set A, 5 on B, 4 on C, 4 on D, 5 on E, 4 on F, and finally, 4 on G. The test ear was chosen by the test person.

2.2.3.3.1 Threshold and slope for words

The students were instructed to use the scoring protocol in Appendix C and tracks 2 to 19, which contained lists of 10 five-word sentences without background noise. The measurement procedure was to start at such a high level that more than 80 % of the words were identified for the first two lists. This guaranteed that the test persons obtained some familiarity with the words. The level was reduced by 5dB after each list until the score dropped below 20 %. Then the level was to be increased in 1 dB steps until the score was again over 80 % and this part of the test was finished. 34 normal-hearing listeners participated in this test.

The results were evaluated by first determining the individual threshold for five-word sentences for each test person by fitting a logistic function to the results obtained on tracks 2 to 19. This five-word sentence threshold were used as a reference for all measurements on the test person, so all the results in the second field test were expressed in sensation level (dB SL) relative to the threshold for five-word sentences. Next, all the measurements were pooled together in 1 dB bins, allowing the responses for each word at different sensation levels to be used to estimate a logistic function for each word. The logistic functions were estimated with the method of least squares by using the solver which is a standard add-in in Excel. The procedure was almost the same as the one described in 2.2.3.2.1, but now the scale was in sensation levels, and the least square errors were weighted with the number of measurements in each bin.

2.2.3.3.2 Verification of the first field test

The 10 five-word sentences from track 35 were mixed with noise with a signal-to-noise ratio of -5.36 dB. This list was tested according to instructions at a comfortable listening level, allowing comparison with the results obtained in the first field test. Identical sentences had been used twice in the first field test, with a signal-to-noise ratio of -4 dB: first in the same order as presented here as sentences 3-12, and later in a different order at the end of the test as sentences 73-82. In the first field test these two repetitions were used to evaluate the training effect. In this second field test the sentences are repeated at the signal-to-noise ratio found to give a 50 % score according to the results from first field test, described in section 2.3.2.1.2.

The scores obtained with this list will be evaluated in order to check whether 50 % recognition is still achieved after the adjustments to the material made after the first field test, as described in 2.2.2.9. We will compare the results from this field test with the results obtained after training from the first field test (sentences 73-82). 107 subjects participated in these measurements in the first field test, while the second field test involved 32 participating subjects.

2.2.3.4 First laboratory test: speech recognition threshold measured in hearing level

This test was performed by students from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College in the department's laboratory during September 2007. The laboratory test incorporated test items from the Quist-Hanssen speech audiometry (results not presented here), three-word utterances (results in Chapter 3), monosyllabic words (results in Chapter 4), five-word sentences (results presented in 2.3.4) and three-numerals lists (results presented in 6.1.1). 34 subjects participated in this test, results are presented from 39 normal hearing ears in a subgroup consisting of 20 persons (15 female and 5 male) with an age between 20-25 years.

The main objectives of testing the five-word sentences were to check the hearing threshold without background noise and to obtain experience using a earphone binaural test consisting of five-word sentences and various types of masking noises. Since this was an in-house test, all the speech audiometers were calibrated before the testing. The results can therefore be presented in hearing level (dB HL), unlike the results in the preceding section, which were expressed in sensation level (dB SL).

An example of the scoring protocol can be found in Appendix D. Five test sets (A, B, C, D and E) were made of the protocol, with a corresponding CD. The structure of these tests was identical, but the sentences and words differed.

The students were instructed to use the scoring protocol in Appendix D, which included two tests using five-word sentences (cf. page 6 and page 8 in the scoring protocol). Page 6 was used for testing tracks 20 and 21, which both comprised a test type called quick-speed test. Each of the tracks contains a list of 20 sentences where each new sentence is reduced by 2.5 dB relative to the one preceding it. The starting level for these tests was obtained by using the pure-tone-average hearing level given in the table on page 7 of the scoring protocol. The binaural test for earphones presented on page 8 of the scoring protocol consisted of nine subtests on tracks 22-30. Each track was a quick-speed list of 10 five-word sentences where each new sentence was reduced by 2.5 dB relative to the one preceding it. The design of the binaural test will be described in section 6.1.3.7.

The results from the quick-speed test on page 4 of the protocol were used to calculate the speech recognition threshold for each of the 39 ears measured, and the median of these thresholds was chosen as the speech recognition threshold for the five-word sentences.

The results of each of the nine subtests in the binaural test on page 8 of the protocol were used to calculate the mean value and standard deviation of the scores for the 14 young normal hearing subjects who participated.

For this test a calibration signal which was 1 dB lower than the equivalent levels of the speech material measured without weighting filter had been calibrated to 20 dB SPL in the earphones of the audiometer. As will be described in section 2.2.3.6, the calibration level of the final material has been altered after this test.

2.2.3.5 Second laboratory test: threshold and slope in noise

This test was performed by the author on students and staff members from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, in the laboratory of the department during April 2008. The laboratory test incorporated five-word sentences, three-word utterances and monosyllabic words. The measurements were performed on one ear for each test subject, using a clinical audiometer calibrated for HiST speech audiometry. Noise from the speech audiometry CD was mixed ipsilaterally with the speech signals. Nine normal hearing subjects participated in the test.

The five-word sentences were presented in lists of 10 sentences. After play-back of one list without noise, the remaining lists were presented at the signal-to-noise ratios of 10, 0, -4, -8 and -12 dB SNR, and the number of correctly recognized words was registered for each sentence. For each subject a different list was selected as the first one, and then the remaining list followed in the same order as the tracks on the CD.

A sigmoid function was fitted to each subject's response by the least squares method using the solution solver in Excel. The mean values and standard deviations of the slope and threshold are presented in 2.3.5.

2.2.3.6 Speech and calibration levels during the listening tests

During the development process of the speech audiometry material the levels of the words and sentences have been measured directly on the wave files with software like Adobe Audition and routines made in Matlab. The levels have been measured as equivalent levels without frequency weighting relative to the 16 bit dynamic range on CDs. The reference signal, a full range square wave has a level of 0dBFSSQ (dB Full Scale SQuare wave). For speech the equivalent levels have been measured with the pauses removed. The levels of the final version of the speech audiometry set have also been measured in an acoustic coupler as will be reported in section 6.1.2.

2.2.3.6.1 First field test calibration

The speech, the noise and the calibration signal (1000 Hz sine) all had the same level at -22.5 dBFSSQ. The calibration of the audiometer did not influence the results, because for all measurements the level was adjusted to a comfortable listening level for the test subject at the start of the session. All the results were based on speech in noise tests were the signal-to-noise ratios were preset on the CD.

2.2.3.6.2 Second field test calibration

The speech level was -42.5 dBFSSQ, the calibration signal was a 1000 Hz sine at -6.8 dBFSSQ the same level as is used on the CDs with Quist-Hanssen speech audiometry.

The students performing the test were instructed to report the VU-level of the calibration signal on the audiometer. These measurement were not influenced by the calibration of the audiometer, because individual reference levels for the threshold of five-word sentences were established during the first measurement. These reference levels were used to convert the results for the test person to dB SL (sensation level) for all the tests performed. The speech level was set so low on the CD to make possible measurements below the hearing threshold also for best scoring listeners.

2.2.3.6.3 First laboratory test calibration

The speech level was -22.5 dBFSSQ and the calibration signal a 1000 Hz 1/3-octave noise had a level of -23.5 dBFSSQ. All the audiometers were calibrated.

2.2.3.6.4 Second laboratory test and final speech audiometry material calibration

The speech level was -25.2 dBFSSQ and the calibration signal a 1000 Hz 1/3-octave noise had a level of -28.5 dBFSSQ. A VU-adjustment tone, a 1000 Hz sine at a level of -18.5 dBFSSQ was introduced, so the strongest parts of the speech material should not overload the audiometer. An extra calibration signal (1000 Hz sine at a level of -8.5 dBFSSQ) for audiometers calibrated to use Quist-Hanssen speech audiometry was included.

The speech level was reduced to avoid that the Quist-Hanssen extra calibration signal was too strong to achieve correct calibration for some combinations of audiometers and CD-players.

The new values for the speech level and the calibration level established 0 dB HL as the speech recognition threshold for three word sentences, based on the results from the first laboratory test.

2.3 Results

2.3.1 Pilot test

2.3.1.1 Naturalness

Table 2.6 shows the results of the pair comparisons for naturalness. The significant results from this evaluation of naturalness have been given a greyscale background shadow. More pronounced results have a darker shadow.

Table 2.7 was prepared in order to clarify whether there were differences between the different level adjustment procedures. In this table, we see for example that xno vs. xuj is 52%, resulting from comparisons of hno vs. huj, wno vs. wuj and dno vs. duj. None of the results in this table are significant, which allows us to concentrate on the comparisons between the different methods for generating Hagerman sentences in the next table.

Table 2.8 presents the level adjustments for each method of making sentences organized as groups. For example, the comparison of dxx vs. hxx is 87%, resulting from comparisons of duj-huj, dno-huj, d52-huj, duj-hno, dno-hno, d52-hno, duj-h52, dno-h52 and d52-h52. Significant results from this evaluation of naturalness have been given backgrounds in different shades of grey. This table gives a clear ranking for naturalness of the speech material made by the different methods. This can be summarized thus:

1. Naturally read sentences (ppp) are evaluated as better than all the other material.
2. Diphone method sentences (dxx) are evaluated as better than Wagener and original Hagerman sentences.
3. Wagener method sentences (wxx) are evaluated as better than original Hagerman sentences.
4. Original Hagerman sentences (hxx) are evaluated as the least natural-sounding material.

Table 2.6 Results in percentage for all the pair comparisons for naturalness. Cells that are positioned symmetrically along the blank diagonal have a total score of 100%. The significant results from this evaluation of naturalness have been given a greyscale background shadow. More pronounced results have darker shadow.

		Evaluated as most natural in pair comparison									
		huj	hno	h52	wuj	wno	w52	duj	dno	d52	ppp
Evaluated as least natural in pair comparison	huj		50	43	79	93	57	71	100	93	86
	hno	50		57	79	71	79	93	93	86	86
	h52	57	43		79	100	64	86	71	93	86
	wuj	21	21	21		64	57	36	86	79	64
	wno	7	29	0	36		43	71	43	79	100
	w52	43	21	36	43	57		57	64	64	71
	duj	29	7	14	64	29	43		43	43	43
	dno	0	7	29	14	57	36	57		50	100
	d52	7	14	7	21	21	36	57	50		50
	ppp	14	14	14	36	0	29	57	0	50	

Table 2.7 Simplified version of Table 4 where only the differences between level adjustment procedures are shown in per cent.

	Evaluated as most natural in pair comparison			
		xuj	xno	x52
Evaluated as least natural in pair comparison	xuj		52	48
	xno	48		50
	x52	52	50	

Table 2.8 Simplified version of Table 4 where the details of the level adjustment procedures are removed, so that only the differences between naturally read sentences (ppp), Diphone sentences (dxx), Wagener sentences (wxx) and Hagerman sentences (hxx) are shown in per cent.

	Evaluated as most natural in pair comparison				
Evaluated as least natural in pair comparison		hxx	wxx	dxx	ppp
	hxx		78	87	86
	wxx	22		64	79
	dxx	13	36		64
	ppp	14	21	36	

2.3.1.2 Speech recognition in noise for material produced by different methods

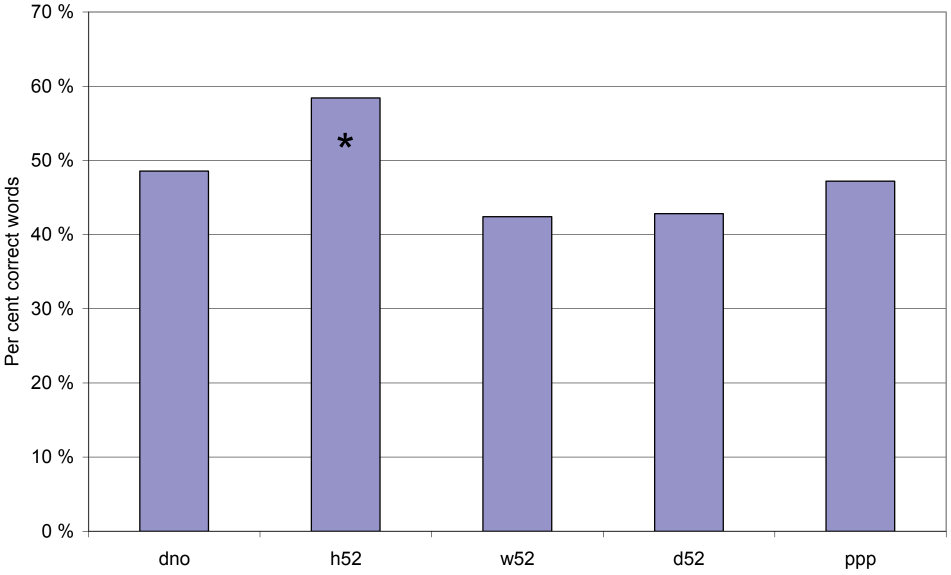


Figure 2.9 Per cent of correctly recognized words for the recognition in noise at -5 dB signal-to-noise ratio for sentences generated by 5 different methods. The group that is significantly different from the others is marked with *.

Figure 2.9 shows the results of the speech recognition in noise test with materials generated by different methods at -5 dB signal-to-noise ratio. Group h52, marked with an asterisk (*), is significantly different from the other groups. Sentences made by using the original Hagerman method are easier to understand at the -5dB signal-to-noise ratio. We find no significant differences between naturally read sentences, sentences made by means of the Wagener method and sentences made by means of the Diphone method. This shows that the fine phonetic details of the natural sentences ppp have not been robust enough to achieve a higher score than the other methods. An explanation for this could be that the noise is so strong that it masks the details at the -5 dB signal-to-noise ratio. The fact that we measure better speech recognition with sentences made by using the original Hagerman method is not an argument against choosing sentences made by means of the Diphone method for speech audiometry. This will only cause a slight shift in the signal-to-noise ratio associated with the speech recognition threshold.

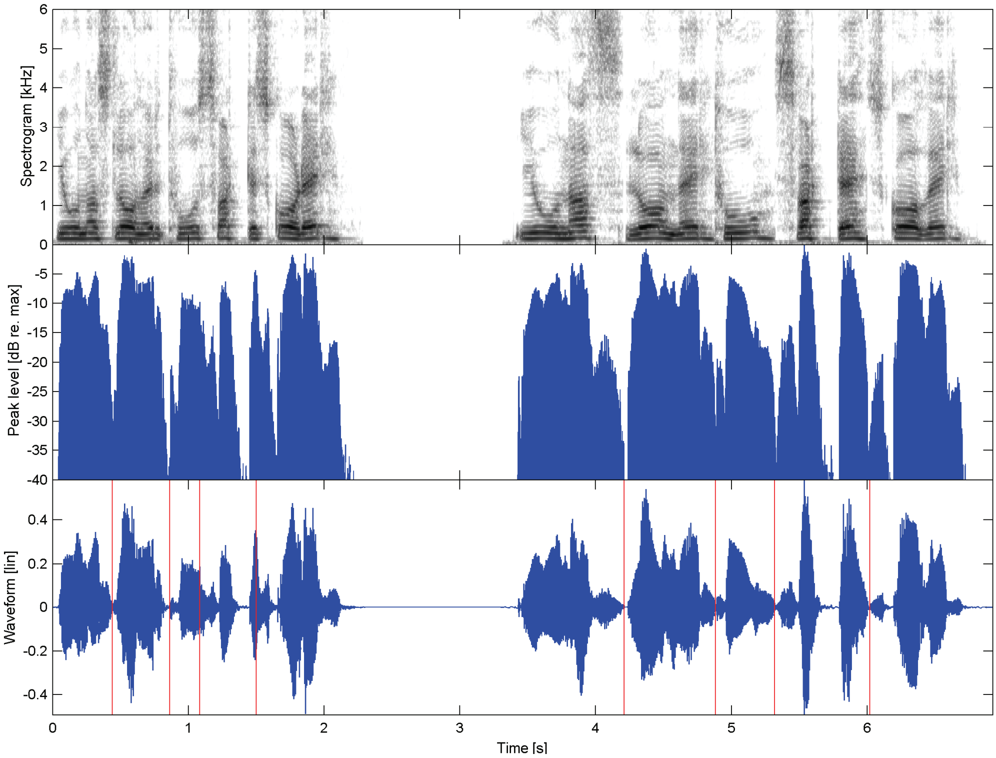


Figure 2.10 Spectrogram, peak levels and waveform of two examples of the sentence "Jonas låner to svarte skåler". The first example is a naturally read sentence, and this is followed by the same sentence made according to Hagerman's original method. The vertical lines in the bottom panel are inserted at the word boundaries.

The reasons for the higher scores for the original Hagerman sentences could be those mentioned in the experimental hypotheses: the staccato sounding original Hagerman material has some qualities, such as more pronounced word boundaries and a slower speech rate, which would be likely to produce a higher score for that material. The original Hagerman sentences are on average 39 % longer than the sentences made with the other methods, which all have approximately the same mean length. Figure 2.10 can help us explain the reasons why the Hagerman method produced better scores.

The figure shows two examples of the sentence "Jonas låner to svarte skåler"; first as a naturally read sentence, and then, after a one-second interval, the same sentence made according to the Hagerman method. The figure shows that the Hagerman sentence is much longer than the naturally read sentence. Some word boundaries are a little more pronounced, since there is a longer pause between words in the Hagerman sentence than in the others. Word durations are also longer, which is likely to make it easier to discern the phonetic characteristics of each phoneme of the word. The characteristics of the present Hagerman material are comparable with those found for clear speech. Picheny et al. (1986) describe some results of acoustic analyses comparing clear speech with conversational speech. The following features are among the ones that are more marked in clear speech: First, speaking rate decreases. This decrease is achieved both by inserting pauses between words and by lengthening the duration of speech sounds. Second, there are differences with respect to the number and types of phonological phenomena such as modification of vowels and release of consonants. Third, the RMS intensities for obstruent sounds are greater. Finally, there are small changes in the long-term speech spectrum. Our sentences made by means of the original Hagerman method have qualities that are similar to those described by the first characteristic identified by Picheny et al. Phonological phenomena and intensity were not analyzed in our material. According to Picheny et al. changes in the long-term spectrum were not substantial. Krause and Braida (2004), on the other hand, found changes in the long-term spectrum to be one of two global-level properties that appear to be linked to the improvements in intelligibility provided by clear speech. Those changes involved, first, increased energy in the 1000-3000Hz range of long-term spectra, and second, increased modulation depth for low frequency modulations of the intensity envelope. Figure 2.11 was made to check the speech spectrum of our material. It shows the third octave spectra of two 10-sentence lists. The lists contain all the words used, so all Hagerman lists will have the same spectrum as the one found on these two. Picheny et al. speculated that the small changes they noticed in the long-term speech spectrum were of little importance for clear speech. We see, however, that the spectrum of our speech material made by means of the

Hagerman method is about 1-2.5 dB higher than the spectrum for our naturally read sentences in the frequency range of 630-3000Hz. This is a very important frequency range for speech understanding and may be a significant factor in the higher intelligibility scores obtained for this material.

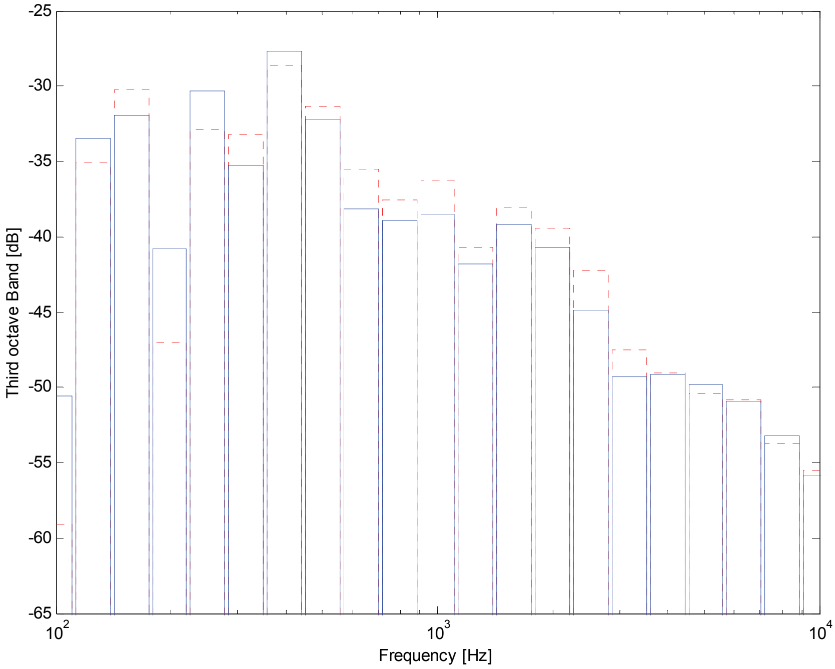


Figure 2.11 Third-octave spectra of two 10-sentences lists. Solid line: naturally read sentences. Dotted line: sentences made by means of the original Hagerman method.

2.3.2 First field test

2.3.2.1 Word score

Excel was used to process the results of the listening tests according to the methods outlined in 2.2.3.2.1. The scores for each word are presented in Appendix B.

2.3.2.1.1 Individual words

Figures 2.12-2.16 on the following pages show the results of the fitted logistic functions for each word as a function of the signal-to-noise ratio. Figures 2.17 and 2.18 presents the thresholds and slopes used to fit the logistic functions. The mean SRT value across all words is -5.36 dB SNR, and the standard deviation 2.13 dB. However, some words have large differences from the mean. *Seks*, the Norwegian numeral six, is the most easily recognisable word with an SRT that is 6.4 dB better than the mean value (Figure 2.14). In contrast, the name *Ingvild* is the least easily recognisable word with an SRT that is 4.3 dB poorer than the mean (Figure 2.12).

Likewise, we also find a great spread in the slopes of the logistic functions. The steepest slope, of 175 %/dB, is found for the word *penner* (pens, Figure 2.16). The smallest slope, of only 6.5 %/dB, is found for the name *Ingvild*, which also had the poorest SRT (Figure 2.12). Because of the two outliers with slopes of over 100 %/dB (Figure 2.18: *Magnus* and *penner*), the mean value of 18.8 %/dB with a standard deviation of 26.0 %/dB for the slopes does not describe this data too well. It may therefore be better to use the median, of 13.3 %/dB, for this purpose.

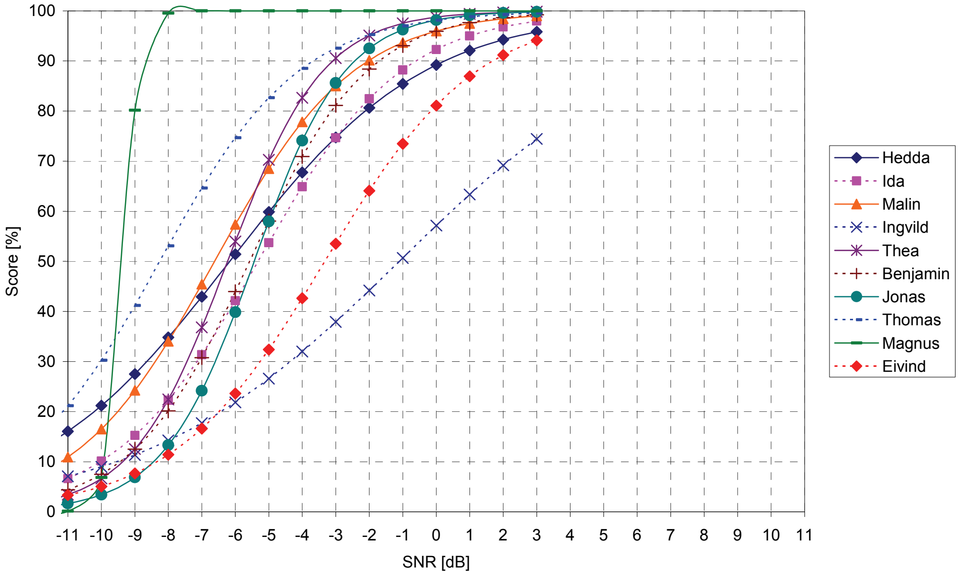


Figure 2.12 The fitted logistic function for the names as a function of signal-to-noise ratio.

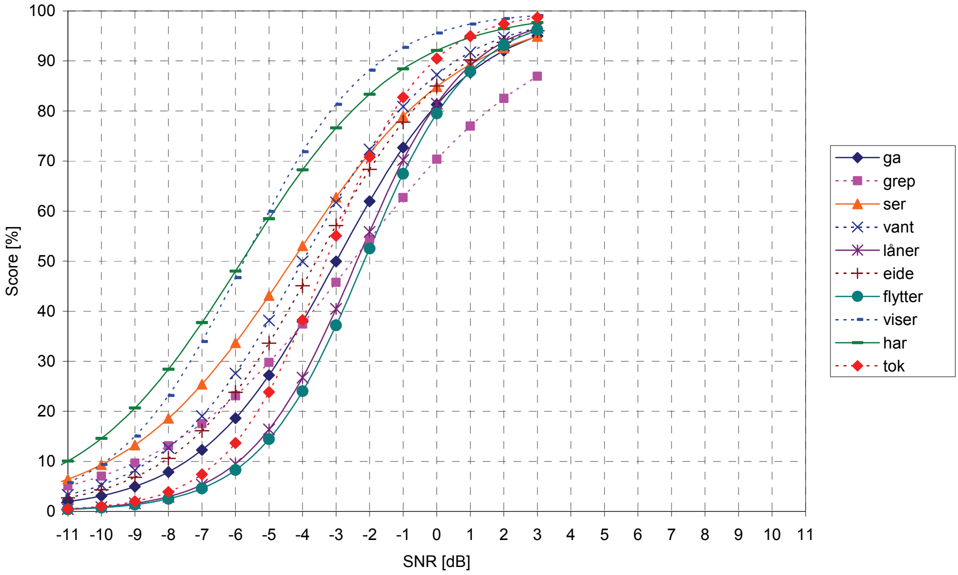


Figure 2.13 The fitted logistic function for the verbs as a function of signal-to-noise ratio.

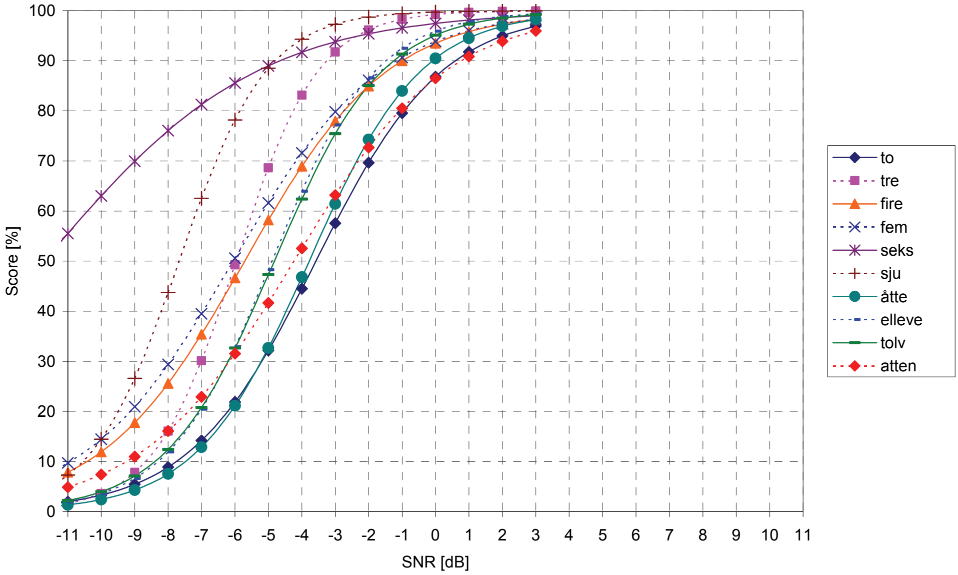


Figure 2.14 The fitted logistic function for the numerals as a function of signal-to-noise ratio.

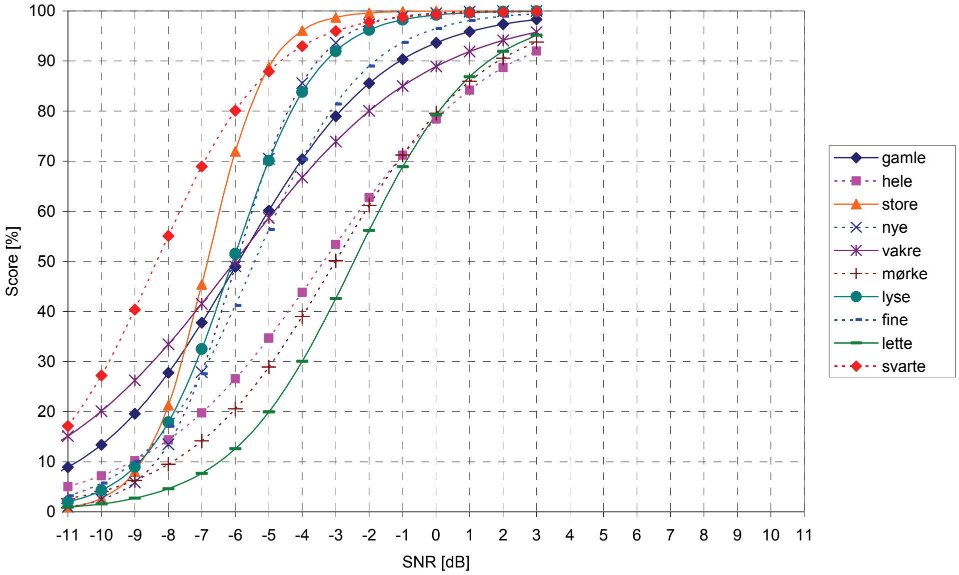


Figure 2.15 The fitted logistic function for the adjectives as a function of signal-to-noise ratio.

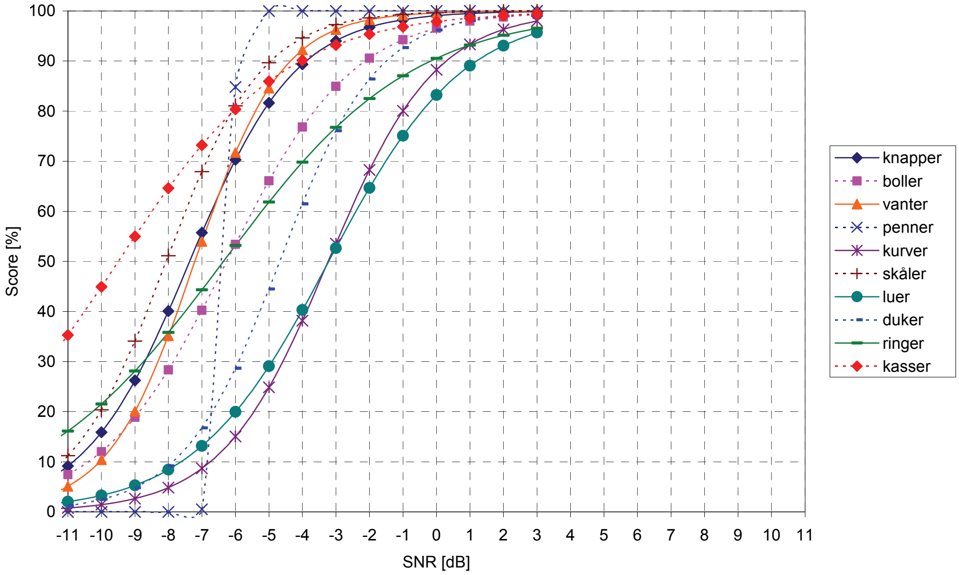


Figure 2.16 The fitted logistic function for the nouns as a function of signal-to-noise ratio.

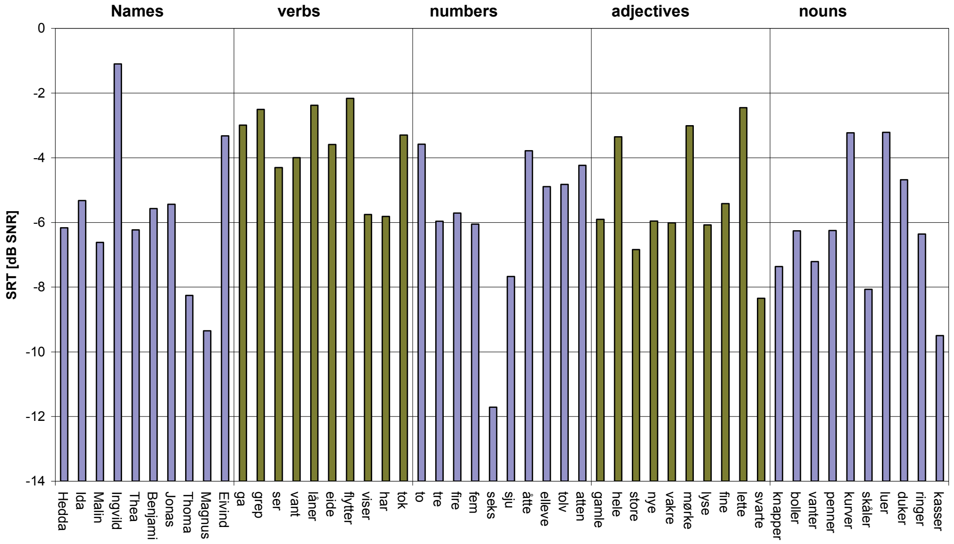


Figure 2.17 The speech recognition threshold for all the words.

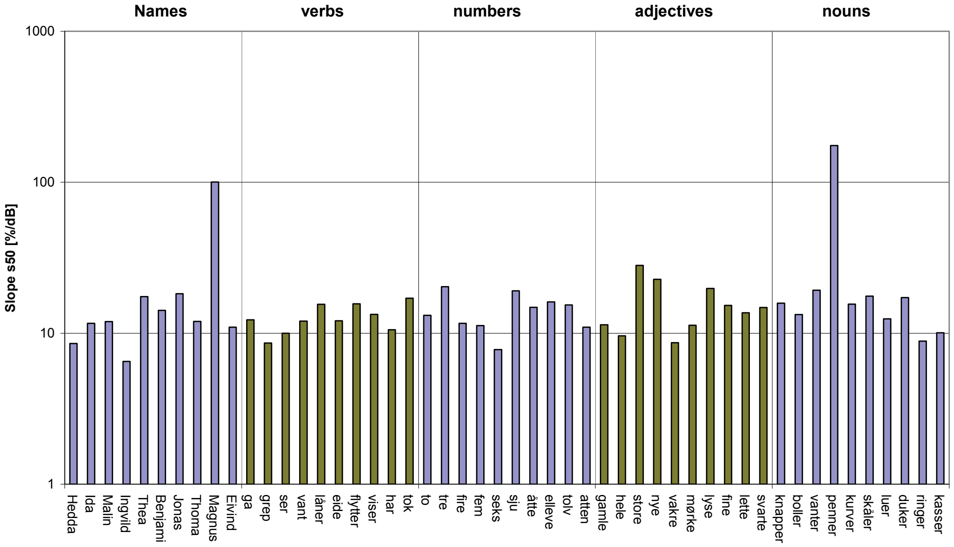


Figure 2.18 The slope in per cent per dB for all the words.

2.3.2.1.2 Word groups

The fitted logistic curves for words of the categories name, numeral and adjective are very similar when we look at the combined data for all words in each group in Figure 2.19. The respective thresholds for the fitted logistic functions are -5.81, -5.58 and -5.32 dB SNR. The nouns have a slightly better threshold at -6.23 dB SNR and their fitted logistic curve is shifted slightly more to the left. In contrast to these values, the verbs have a markedly poorer threshold of -3.65 dB SNR, and the fitted logistic function of the verbs is shifted substantially to the right; see Figure 2.19.

The slopes are relatively similar for all the word groups, with a range of 10.1 to 11.6 %/dB for the fitted logistic functions; see Figure 2.19.

Likewise, the speech recognition threshold of -5.31 dB SNR and slope of 10.6 %/dB for the logistic function fitted to all the words (see Figure 2.20), are as expected from the data for the word groups. We expect the sentence scores to have the same values for these parameters as the values found here for the total of all the words. There is a small difference between this SRT of -5.31 dB SNR for one logistic function fitted to all the words, and the value of -5.36 dB SNR found in section 2.3.2.1.1 as the mean SRT for all of the logistic functions fitted to the individual words. This insignificant difference probably arises from approximations performed when fitting the logistic functions with the method of least squares.

However, a value of 10.6 %/dB for the slope of the sentences is a poor score compared to the median slope for the words, which was found to be 13.3 %/dB in the preceding section. We can implement the probabilistic model proposed by Kollmeier (see section 2.2.3.2.1) and compute the estimated $s50_{sentence}$ value using Equation (2.2).

If we use the mean slope $s50_{word}$ value of 18.8 %/dB and the standard deviation of 2.13 dB for the thresholds of the individual words from the preceding section, Equation (2.2) gives us the estimated slopes for the sentences as $s50_{sentence}$=13.2 %/dB. This is far from our value of 10.6 %/dB. However, as we stated in the preceding section the mean values for word slope are poorly suited as a description of the data because of some of the outliers, and we proposed to use the median instead. If in Equation (2.2) we exchange the mean value of the slope for words with the median value of 13.3 %/dB, we get an estimated value of the sentence slope of $s50_{sentence}$=10.8 %/dB, which corresponds very well with the value of 10.6 %/dB found for all the words in Figure 2.20.

A steep slope is a prerequisite for obtaining reliable speech audiometry measurements in the shortest possible time. Equation (2.2) shows that if we can reduce the standard deviation of the thresholds for the individual words, we achieve a steeper slope for the sentences ($S50_{sentence}$). Reduction of the standard deviations of the thresholds for the individual words can be obtained by adjusting the level of the words in order to make the less easily recognizable words easier to recognize; and vice versa for the words that are recognized more easily. The results of this field test were used to introduce such level adjustments of the individual words. The level adjustment procedure was described in section 2.2.2.9.1.

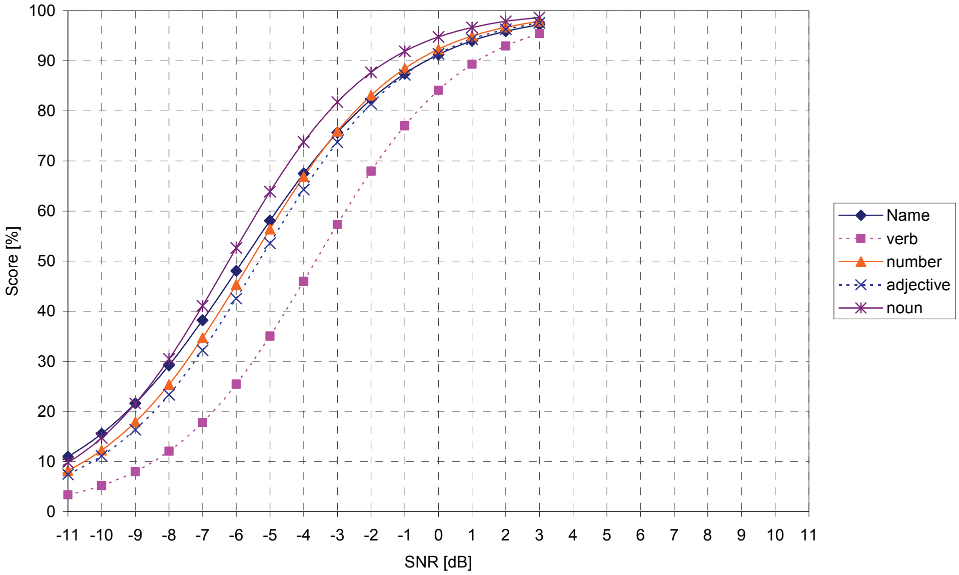


Figure 2.19 The fitted logistic function for the word groups as a function of signal-to-noise ratio.

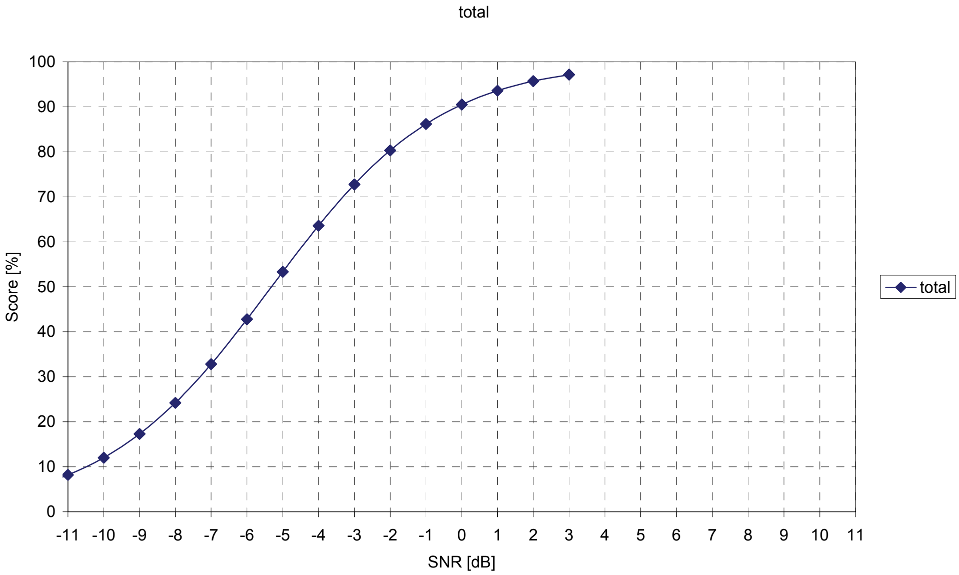


Figure 2.20 The fitted logistic function for all the words as a function of signal-to-noise ratio.

2.3.2.2 Training effect

The test protocol included two lists of the same 10 sentences repeated at -4 dB SNR. The sentences numbered 3-12 were presented first, and repeated last, but in a different order as sentences 73-82. The mean score for the initial -4dB SNR test list amounted to 37.6 % for all the test persons, whereas the mean score for the final test list, containing the same sentences in a different order, amounted to 65.3 %. If we use this data with the fitted logistic function for all words in Figure 2.20, we find that this corresponds to a 2.7 dB improvement of the threshold. This training effect is greater than the one found by Wagener, which corresponded to 2.2 dB. A reason for the greater training effect with our data may be that our listeners had heard identical sentences once before when the sentences was repeated in the final list. The reason for comparing identical sentences was to reduce the influence from possible differences in thresholds for different lists, and believing that repeating sentences will have a small influence on the results when they have a different order and the test person have listened to 60 other sentences in between. Another reason for the differences between the results can be a result from the design where Wagener used lists of 20 sentences and we used lists of 10 sentences. Starting the test with a list of 20 sentences gives more training within the list initially, which can explain some of the difference between the results.

When designing measurements it is important to be aware of this training effect, but if the test persons are familiarised with the material initially through about 20 sentences, the training effect should be no greater than some 1 dB.

2.3.2.3 Dialects

Figures 2.21 to 2.25 show the scores for all the words among each dialect group for sentences 73-82 with a signal-to-noise ratio of -4 dB. If any word was particularly easy or difficult to recognize for a specific dialect group we ought to be able to detect this visually in these graphs. We need to keep in mind that the measurement data for persons from the Central dialect group is based on only two individuals. The results from this group should therefore be downplayed. The visual impression is that there are no great differences between the dialect groups: There are some variations but for the most part we see that there is good agreement.

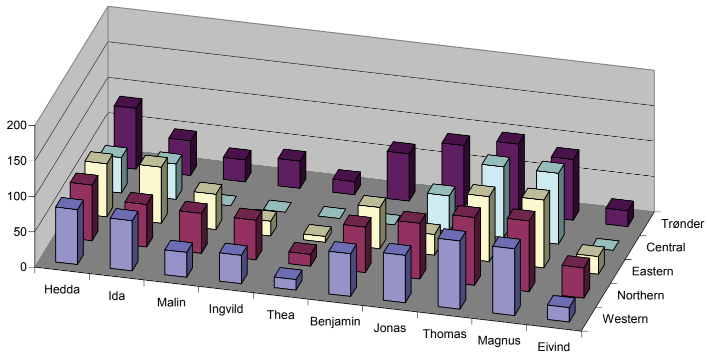


Figure 2.21 Score for names in the dialect groups.

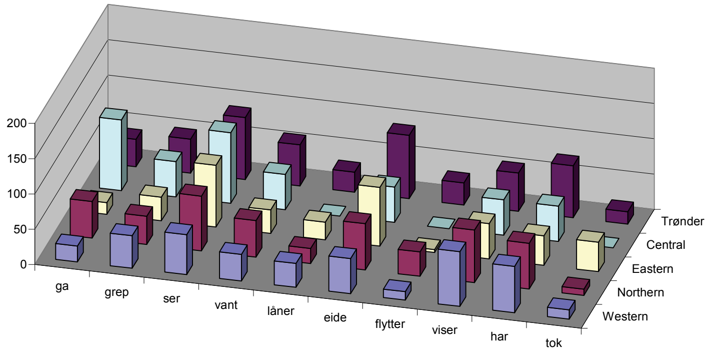


Figure 2.22 Score for verbs in the dialect groups.

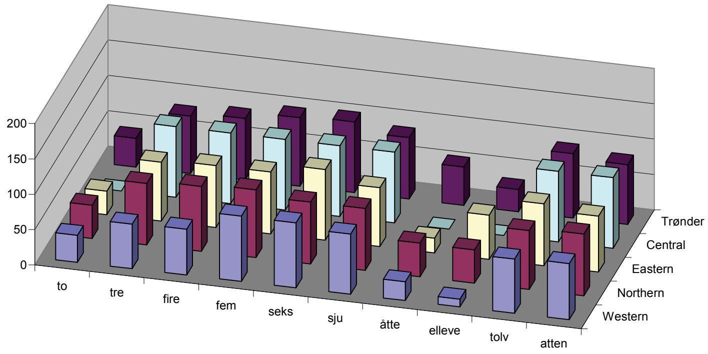


Figure 2.23 Score for numerals in the dialect groups.

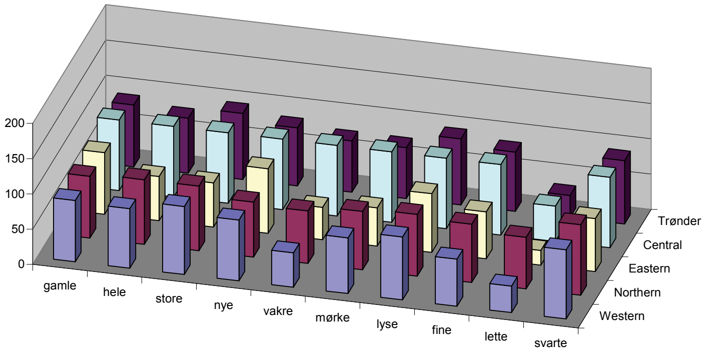


Figure 2.24 Score for adjectives in the dialect groups.

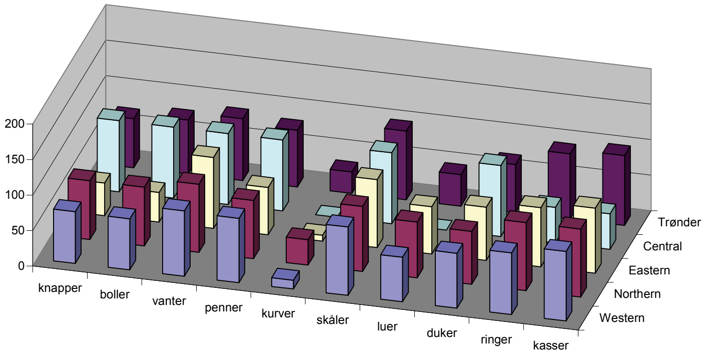


Figure 2.25 Score for nouns in the dialect groups.

Table 2.9 The correlation between the dialect groups for the 50 word scores measured with a signal-to-noise ratio of -4 dB.

Dialect group	Western	Northern	Eastern	Central	Trønder
Correlation with all	0.97	0.94	0.84	0.81	0.95
Correlation with Western		0.90	0.80	0.75	0.88
Correlation with Northern			0.76	0.78	0.88
Correlation with Eastern				0.64	0.80
Correlation with Central					0.78

We can also check for score differences between the dialect groups by computing the correlation coefficient for each group compared to the others. The results are given in Table 2.9. If we exclude the correlations involving the Central dialect group, which only contained two test persons, we find the correlations between the dialect groups to range from 0.76 to 0.90. The correlations were highly significant for all the dialect groups, at n=50, p<0.001.

Even if the correlation was good there may be significant differences between the dialect groups for some of the words. To test this we exclude the central dialect with insufficient data and for each word use a 2x4 contingency table with a column for each dialect and the frequencies of the words recognized in one row and the frequencies of the words not recognized in the other row. Since some of the frequencies are below 5 the chi-square distribution are inadequate to evaluate if the scores for some words have a significant different distribution among the dialects. Cardillo (2008) have made available a Matlab routine for doing Fisher's exact test of 2x4 contingency tables which will give the precise probabilities for the 2x4 contingency table for each word. Table 2.10 on the next page gives the result of using the Fisher's exact test on our data. If we select p<0.05 as our level of significance, we have to remember that we are performing multiple tests with 50 words and we select to use the Bonferroni correction for the level of significance p<0.05/50 which is p<0.001. Even if eight of the words have computed probabilities below 0.05 (marked with a *), none of the computed probabilities are below the level of significance, so we may conclude that we have no sign that there exist significant differences for any of the words between the dialect groups.

Finally, we can evaluate the material by computing the logistic functions for each word group within the dialect groups. We use the measurements made on 40 sentences (33-72) for this analysis; hence, this data is based on 4 times as many measurements as those described earlier in this section (73-82). Figure 2.26 presents the estimated thresholds and Figure 2.27 the estimated slopes from this analysis. The largest threshold difference between the dialect groups is found for the nouns, where Northern dialect background produces a threshold that is 1.8 dB better than the equivalent threshold for subjects with Eastern dialect background. However, the thresholds for each dialect group lie within a range of between -0.7 dB and +0.4 dB discrepancy from the combined threshold for all dialect groups and all the words (TOTAL in Figure 2.26). Likewise, the discrepancy between the slopes for each dialect group and the combined slope for all the dialects are small, with no differences falling outside the range of -12 % (-1.3 %/dB) to +12 % (+1.4 %/dB) for the specific word groups if we exclude the Central dialect group, for which we have insufficient data. This picture improves even further when we look at the combined data for all the word groups (TOTAL in Figure 2.27), where the greatest difference between dialect groups is less than 10 % (1 %/dB), and no dialect group has differences outside a range of -6 % (-0.6 %/dB) to +4 % (+0.4 %/dB) from the combined slope of all the measurements.

Table 2.10 The results of using Fisher's exact test for all the words

	Frequency of recognized words				Frequency of not recognized words				
Word	Western	Northern	Eastern	Trønder	Western	Northern	Eastern	Trønder	two-tailed probability (* - p<0.05)
Hedda	30	18	9	26	9	5	3	5	0.8489
Ida	27	13	9	15	12	10	3	16	0.2345
Malin	13	13	6	9	26	10	6	22	0.1466
Ingvild	15	12	2	11	24	11	10	20	0.2353
Thea	5	3	1	5	34	20	11	26	0.9531
Benjamin	23	14	7	20	16	9	5	11	0.9743
Jonas	26	18	3	27	13	5	9	4	*0.0011
Thomas	37	22	11	30	2	1	1	1	0.8504
Magnus	36	23	11	26	3	0	1	5	0.1984
Eivind	7	9	3	6	32	14	9	25	0.2709
ga	8	11	2	12	31	12	10	19	0.0738
grep	17	9	4	14	22	14	8	17	0.9021
ser	22	17	10	27	17	6	2	4	*0.0301
vant	14	11	4	18	25	12	8	13	0.2577
låner	13	4	3	8	26	19	9	23	0.6000
eide	19	15	10	27	20	8	2	4	*0.0037
flytter	4	7	0	9	35	16	12	22	*0.0298
viser	30	17	6	16	9	6	6	15	0.0778
har	25	14	5	22	14	9	7	9	0.3557
tok	5	2	5	5	34	21	7	26	0.1004
to	14	10	4	12	25	13	8	19	0.9184
tre	24	19	10	25	15	4	2	6	0.1905
fire	25	21	10	26	14	2	2	5	0.0645
fem	36	22	10	30	3	1	2	1	0.4186
seks	35	20	12	31	4	3	0	0	0.1331
sju	33	20	10	26	6	3	2	5	1.0000
åtte	10	10	2	16	29	13	10	15	0.0587
elleve	4	10	7	9	35	13	5	22	*0.0019
tolv	29	19	10	28	10	4	2	3	0.4099
atten	30	20	9	26	9	3	3	5	0.6957
gamle	33	20	10	27	6	3	2	4	1.0000
hele	32	21	7	24	7	2	5	7	0.1435
store	37	21	7	29	2	2	5	2	*0.0093
nye	33	17	11	25	6	6	1	6	0.6216
vakre	19	17	5	22	20	6	7	9	0.0677
mørke	30	19	6	22	9	4	6	9	0.2050
lyse	34	20	10	29	5	3	2	2	0.7173
fine	26	19	8	26	13	4	4	5	0.2675
lette	14	16	2	9	25	7	10	22	*0.0054
svarte	38	23	9	27	1	0	3	4	*0.0177
knapper	28	19	5	21	11	4	7	10	0.1070
boller	27	19	5	23	12	4	7	8	0.1019
vanter	35	22	12	27	4	1	0	4	0.5689
penner	35	18	8	24	4	5	4	7	0.2351
kurver	4	8	1	9	35	15	11	22	0.0519
skåler	37	21	11	30	2	2	1	1	0.6853
luer	24	18	8	13	15	5	4	18	0.0544
duker	30	17	9	20	9	6	3	11	0.7073
ringer	33	22	10	28	6	1	2	3	0.5031
kasser	38	22	11	30	1	1	1	1	0.7301

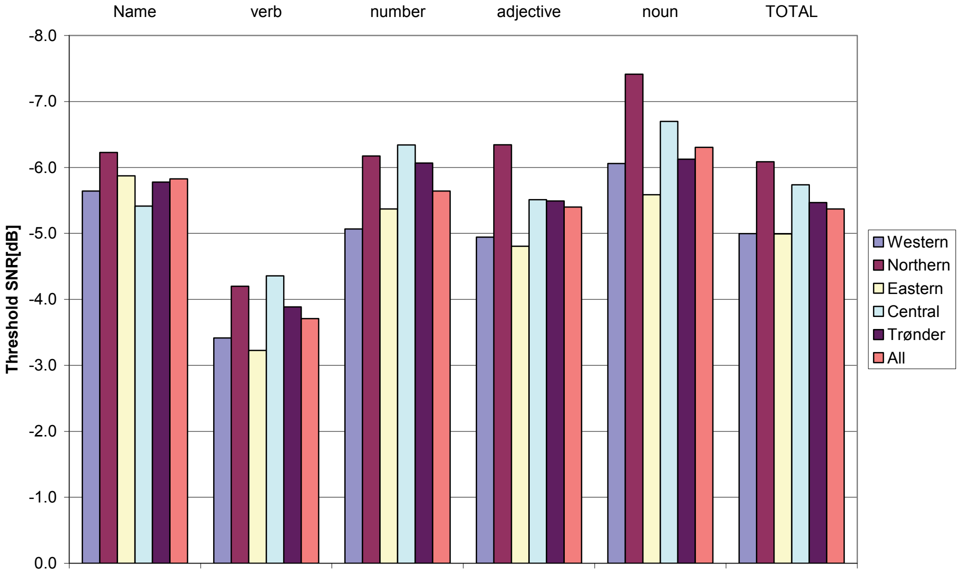


Figure 2.26 The thresholds for logistic functions fitted to each dialect group for the different word groups.

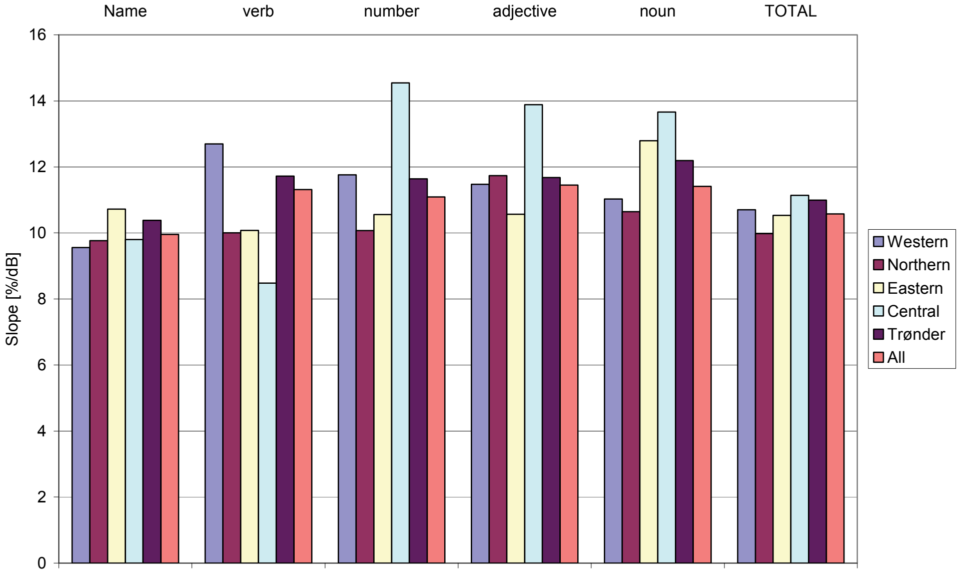


Figure 2.27 Slopes s50 for logistic functions fitted to each dialect group for the different word groups.

To evaluate the influence of Norwegian dialects on speech audiometry measurements, we can summarize first the experiences with dialects and speech audiometry reported by Stensby et al. (2002) (referred to in section 2.2.3.2.3), who found that speakers of an Eastern Norwegian dialect produce the lowest error rates among listeners irrespective of their own dialect. Second, as shown by Figures 2.21-2.25, no substantial difference is found for any one word in any of the dialect groups compared to the other dialect groups in terms of how easily it can be recognized. Third, this is confirmed by the relatively large correlation coefficients in Table 2.9, none of which is smaller than 0.76. It is also not found significant differences for the words between the dialects as presented in Table 2.10. Furthermore, the thresholds for the logistic functions fitted to the dialect groups have no greater difference than -0.7 dB and +0.4 dB from the combined total. This is still the case when we include the scarce data from the Central dialect group. Finally, all the slopes for the logistic functions fitted to the dialect groups lie within a range of -6 % to +4 % from the combined slope. In view of all these arguments, our impression is that speech audiometry materials made with a speaker from the Eastern dialect group will function very well among listeners irrespective of their own dialect.

2.3.2.4 Noise type

A check of whether the temporal fixation between the noise and the speech on the compact disks influenced the score from test sets A-D was performed using the following procedure: The signal-to-noise of all the words was calculated using Matlab and Excel for all the test sets A-D. The differences between these values and the mean signal-to-noise ratios for the four occurrences of the same word in sets A-D were calculated. These calculations produced 200 SNR differences (50 for each test set A-D). Likewise, the differences were calculated between the score for each word on each of the sets (A-D) and the mean score value for the same word on all of the sets A-D combined. These calculations provided 200 score differences. Figure 2.28 shows a scatter plot for all these score differences as a function of the corresponding SNR differences.

If the noise that was temporally fixed to the speech on the compact disks had any influence on the scores, we would expect a positive trend in this plot. Instead, we find that the fitted line is almost horizontal with a small negative slope, but with a very poor fit of r^2=0.014. We can interpret this result primarily as an indication that the use of noise that was temporally fixed to the speech on the compact disks did not have any negative impact on our measurements.

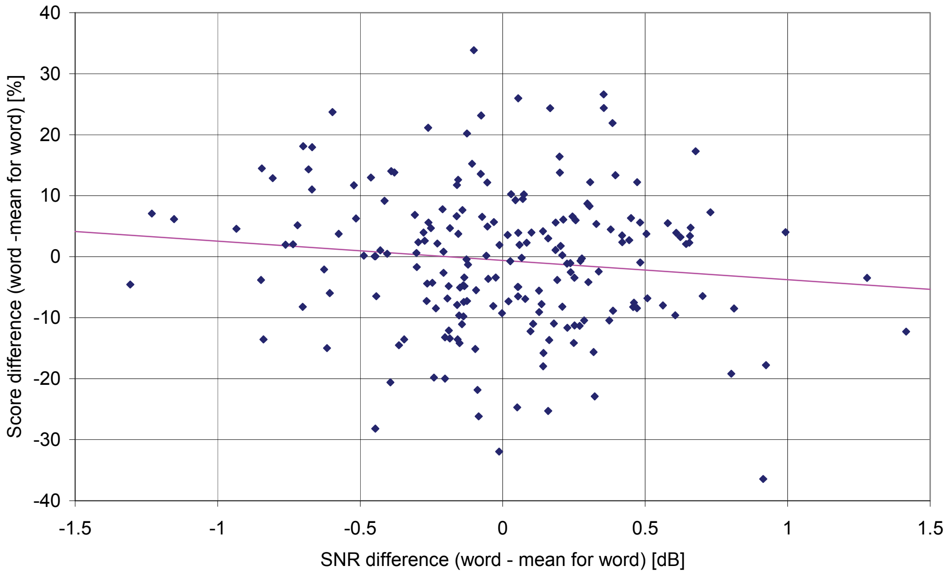


Figure 2.28 The differences between the score for words in test sets A, B, C and D and the mean score for the word as a function of the differences between the SNR for the same words and the mean SNR for the word. The determination coefficient for the fitted line y=-3.16x-0.62 is r^2=0.014.

2.3.3 Second field test

2.3.3.1 Word score

The results from the listening tests were evaluated using Excel following the procedures outlined in 2.2.3.3.1.

2.3.3.1.1 Individual words

Figures 2.29-2.33 on the following pages show the fitted logistic functions for each word as a function of the sensation level. The sensation level was referenced to the speech recognition threshold for five-word sentences for each subject. The mean SRT value across all words is 0.1 dB SL, with a standard deviation of 2.0 dB. The largest differences from the mean are the values for the noun *boller*, which has an SRT 7.0 dB better than the mean; and the noun *kasser*, whose SRT is 4.6 dB below the mean (Figure 2.33).

The mean and median slopes of the words are very similar, at 10.9 %/dB and 10.5 %/dB, respectively. The standard deviation is 2.9 %/dB. The variations are within 5.1 %/dB for the name *kasser* (Figure 2.33) and 17.4 %/dB for the name *Jonas* (Figure 2.29).

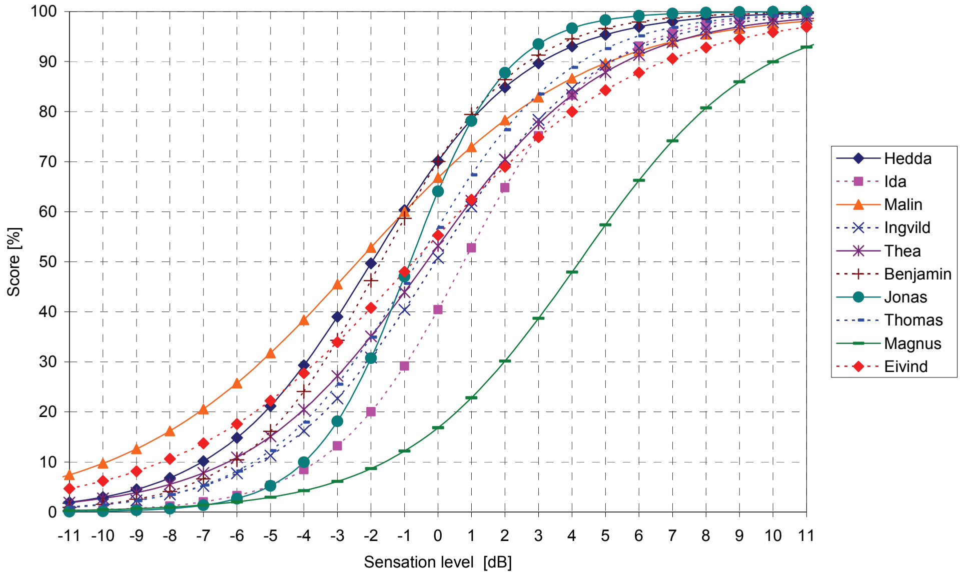


Figure 2.29 The fitted logistic function for the names as a function of sensation level.

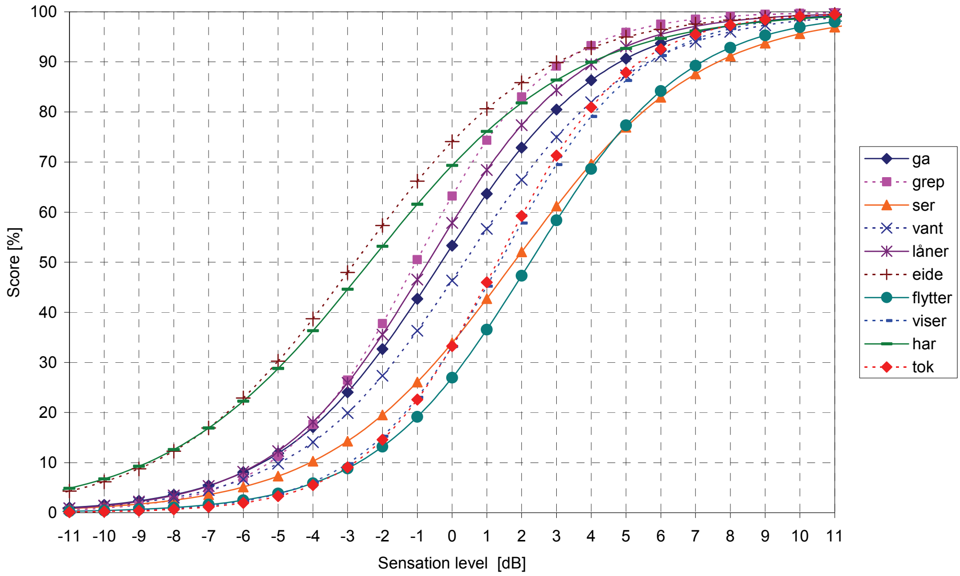


Figure 2.30 The fitted logistic function for the verbs as a function of sensation level.

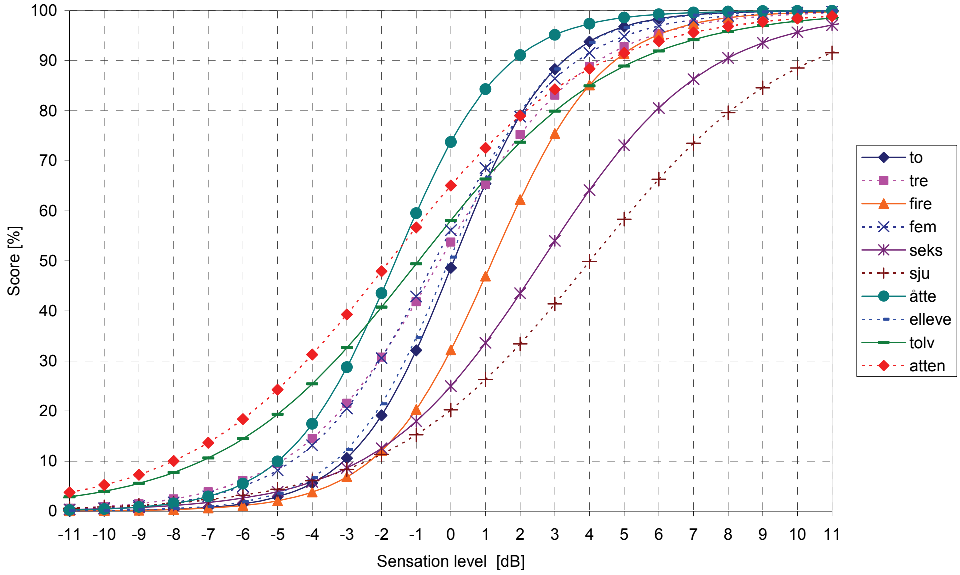


Figure 2.31 The fitted logistic function for the numerals as a function of sensation level.

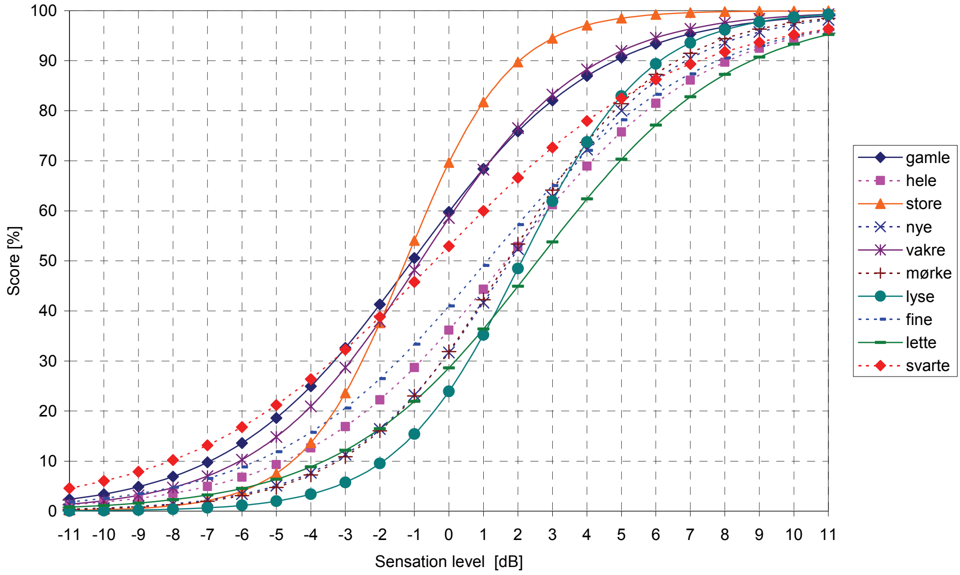


Figure 2.32 The fitted logistic function for the adjectives as a function of sensation level.

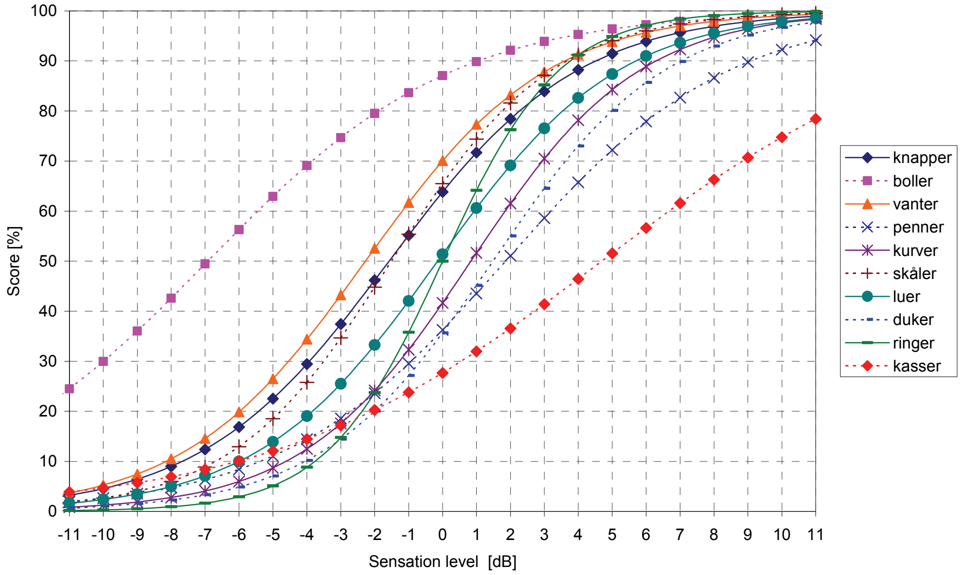


Figure 2.33 The fitted logistic function for the nouns as a function of sensation level.

2.3.3.1.2 Word groups

The fitted logistic curves for names, verbs, numerals, adjectives and nouns (Figure 2.34) are very similar and close to the fitted logistic curve for all the five-word sentence words (Figure 2.35).

The SRTs for names, verbs, numerals, adjectives and nouns differ by respectively -0.4, 0.1, 0.2, 0.7 and -0.4 dB from the SRT for all words. The slopes of names, verbs, numerals, adjectives and nouns are, respectively, 9.5, 10.2, 11.1, 9.5 and 8.1 %/dB; and for all the words the slope of the fitted logistic function is 9.6 %/dB.

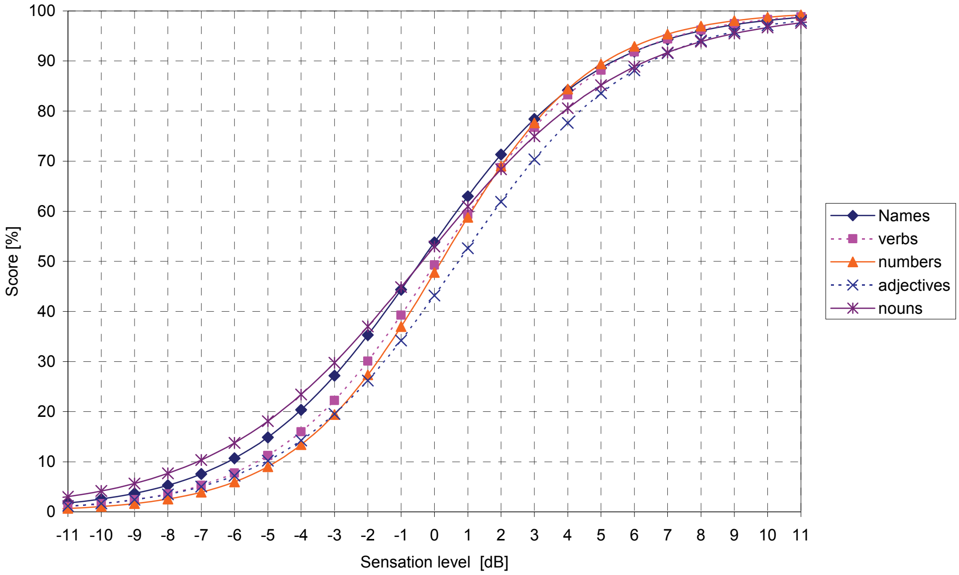


Figure 2.34 The fitted logistic function for the word groups as a function of sensation level.

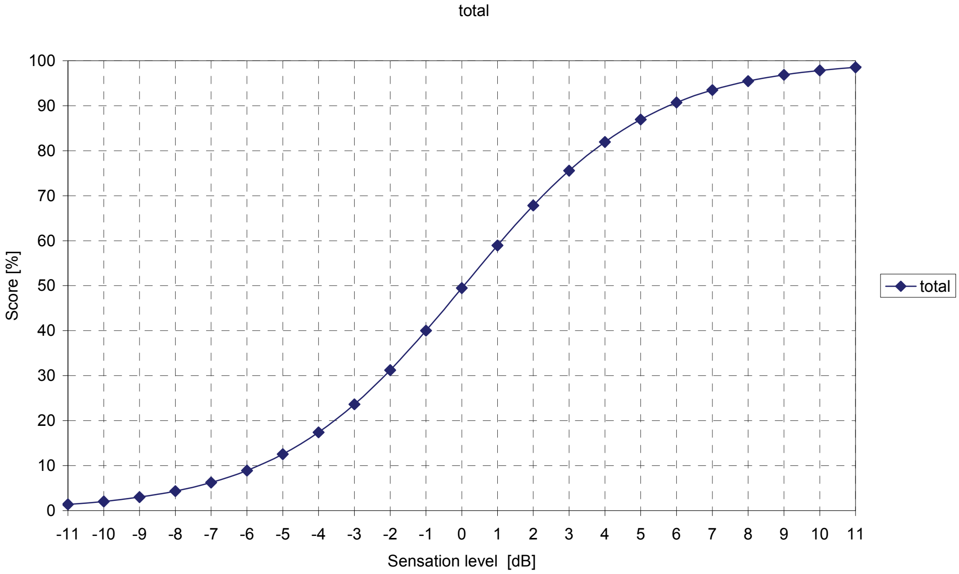


Figure 2.35 The fitted logistic function for all the words as a function of sensation level.

2.3.3.2 Influence of measurement method

The sensation levels used to evaluate responses in the second field test were established by fitting a logistic function for each test person as described in 2.2.3.3.1. This logistic function was fitted by using all the results from the 5 dB descending steps, and the 1 dB rising steps, for sentences 2-19 in the protocol. This method is here called the *full method*. An alternative and quicker measurement procedure would be to use only the results from the 5dB descending steps to estimate the logistic functions. We will call this method the *quick method*, and a histogram of the differences between the thresholds estimated by the two methods is shown in Figure 2.36. We had 25 measurements on which we could perform this comparison, and in 76 % of the cases (19 measurements) the difference was smaller than 1 dB. The mean difference between the two methods was -0.19 dB, with a standard deviation of 1.12 dB.

Likewise, Figure 2.37 shows the slopes of the logistic functions estimated by the two methods. with a perfect match here, all the points would have been located along the diagonal. Hence, there is some discrepancy between the two methods, but no systematic bias.

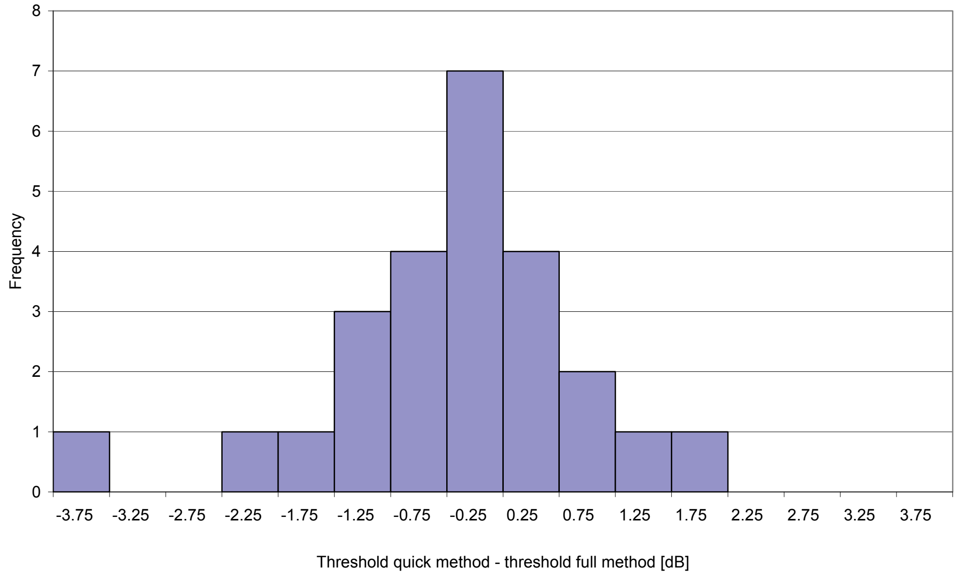


Figure 2.36 Histogram of the difference between the thresholds measured by two methods.

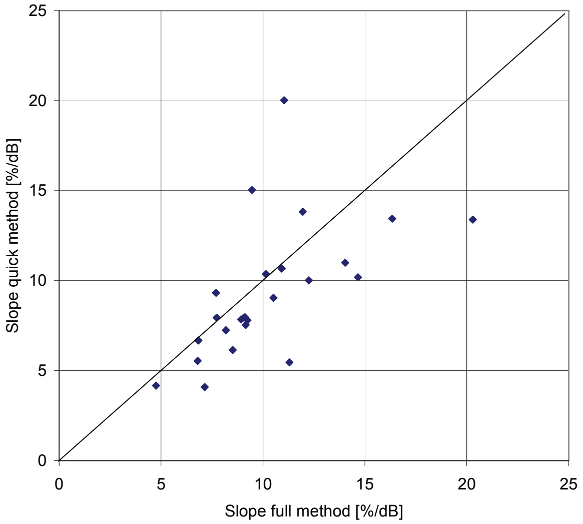


Figure 2.37 The slopes of the quick method as a function of the slopes of the full method.

2.3.3.3 Verification of the first field test

The mean score in the first field test for the list of 10 five-word sentences, with an SNR of -4 dB, is 65.3 %, with a standard deviation of 13.4 %. For the same sentences in the second field test, with an SNR of -5.36 dB, the mean score is 47.7 % and the standard deviation 10.5 %.

If we calculate statistics after excluding the first three sentences for the reasons given in the next paragraph, we get a mean score of 66.2 % for sentence 4 – sentence 10 in the first field test (SNR -4 dB), with a standard deviation of 12.6 %. For the second field test (SNR -5.36 dB), the mean is 53.2 % and the standard deviation 5.2 %.

Inspection of Figure 2.38 shows that scores for sentence 1 – sentence 4 in field test 2 (FT2) seem to increase for each new sentence measured. This behaviour can perhaps be explained by the fact that the listeners were unfamiliar with listening to sentences in noise: Even though this test was performed after the subjects had received substantial training with the test words and sentence structure by listening to about 190 five-word sentences and 150 three-word utterances, they had never listened to these sentences mixed with noise before. The increasing scores may therefore be a result of familiarization with the test situation of listening to sentences mixed with noise.

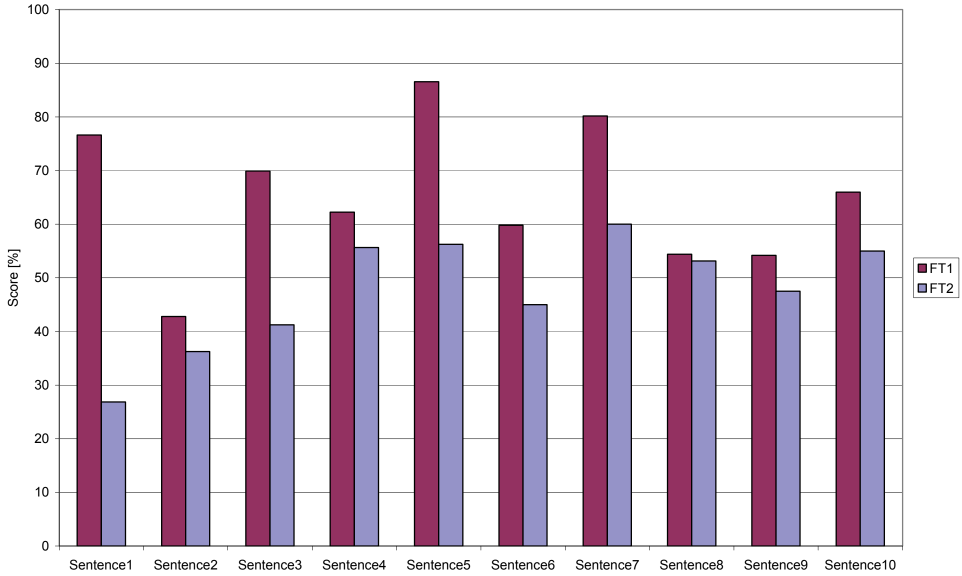


Figure 2.38 The left hand columns show the score of sentences with an SNR of -4 dB from the first field test. The right hand columns show the score of the same sentences with an SNR of -5.36 dB in the second field test.

We were expecting a score close to 50 % in the second field test. The 53.2 % value obtained for the last 7 sentences is very close to this expected value. Hence, we can conclude that the adjustments made to the level of the words after the first field test do not seem to have influenced the expected threshold in noise for these five-word sentences. We can also speculate that the reduction in standard deviation for these 7 sentences from 12.6 % in the first field test to 5.2 % in the second field test may be a result of these level adjustments.

Finally, we can use the slope of 10.6 %/dB found for the five-word sentences with noise from section 2.3.2.1.2 to estimate the thresholds in noise (50 % score) for the seven last sentences (sentence 4 – sentence 10). These calculations give an estimated threshold of -5.53 dB SNR for the first field test and -5.66 dB SNR for the second field test. The very small difference of 0.13 dB between these two estimations confirms the conclusion in the preceding paragraph that the adjustments made to the levels of the words do not seem to have influenced the threshold in noise for these sentences.

2.3.4 First laboratory test: threshold measured in hearing level

Measurements and evaluation of the results were carried out as described in section 2.2.3.4. The median value of the speech recognition thresholds was 3.8 dB HL. This threshold was 1.0 dB higher than what was found to be the case for three-word utterances measured with the same method. The level of the calibration signal has later been altered as described in section 2.2.3.6.

Table 2.11 Mean values and standard deviations of the thresholds for the subtests in the five-word sentences binaural test measured in signal-to-noise ratio dB. The results of subtests 1-3, which are measured with at least one ear without masking noise, cannot be presented as dB SNR. However, the mean values of recognized words were 50.0, 49.6 and 49.9 for subtests 1-3 respectively.

Speech and noise type	mean SRT [dB SNR]	sd [dB]
1. Speech binaural, no noise.		0.0
2. Speech binaural, noise left ear.		0.3
3. Speech binaural, noise right ear.		0.2
4. Speech binaural phase shifted, noise binaural.	-14.3	2.4
5. Speech binaural, noise binaural phase shifted.	-14.3	2.7
6. Speech binaural, noise temporally simulated in left ear by delaying noise 0.6 ms in right ear.	-12.2	1.8
7. Speech binaural, noise temporally simulated in right ear by delaying noise 0.6 ms in left ear.	-11.8	2.4
8. Speech binaural, noise binaural uncorrelated.	-8.8	1.6
9. Speech binaural, noise binaural.	-7.8	1.1

Table 2.11 presents the mean values and standard deviations of the thresholds in the nine subtests of the five-word sentences binaural test. The design of the binaural test will be described in section 6.1.3.7. These results were used to mark threshold ± one standard deviation in the protocol sheets used for this test.

2.3.5 Second laboratory test: threshold and slope in noise

Measurements and evaluation of the results were carried out as described in section 2.2.3.5. The mean value of the threshold was -6.0 dB SNR with a standard deviation of 0.8 dB. The mean value of the slope was 14 %/dB with a standard deviation of 3.4 %/dB.

Wagener (2003) presents data for the Swedish, German and Danish five-word sentences. These data are compared to the Norwegian results in Table 2.12.

Table 2.12 Comparison of slope and speech recognition thresholds in noise across four languages.

Language	slope [%/dB]	SRT [dB SNR]
Swedish	16.0	-8.1
German	17.1	-7.1
Danish	13.2	-8.4
Norwegian	14	-6.0

2.4 Discussion

2.4.1 Pilot listening tests

Based on the data in section 2.3.1.1 we can conclude that sentences made by using the Wagener and Diphone methods are more natural-sounding than the original Hagerman sentences. None of the methods described gives the same naturalness as naturally read sentences, but sentences made by means of the Diphone method are better in this respect than those produced by using the Wagener method.

Diphone sentences are very well suited to the purpose of performing a speech audiometry sentence test. The material has all the qualities of the Hagerman material, where all the lists are perfectly phonemically equalized, allowing a large number of test sessions to be run. The speech is more natural-sounding than both the Hagerman and the Wagener sentences.

When measuring speech recognition in noise the differences are small between the Hagerman sentences made by means of the Wagener and the Diphone methods on the one hand, and naturally read sentences on the other. Interestingly, the original Hagerman sentences – the least natural sounding sentence type – is the only type of sentence material that gets significantly better speech recognition results in noise. The explanation for this better score may be that the material generated by means of the original Hagerman method has some of the qualities described for clear speech. The speech rate is slower, with more marked pauses between words, and the duration of individual speech sounds is lengthened. When comparing the spectra for the original Hagerman sentences with naturally read sentences, we find for the original Hagerman sentences a small increase in the most important frequency range for speech understanding, which may be another reason for the better score.

2.4.2 First field test

The speech-recognition-threshold signal-to-noise ratio and the slope of the performance-intensity curves for all the individual words were estimated from the results of the first field test. The thresholds varied between -6.4 dB and +4.3 dB relative to the mean threshold which was found to be -5.36 dB SNR (section 2.3.2.1.1). The standard deviation of the thresholds was 2.13 dB. The median slope for the words was found to be 13.3 %/dB. The thresholds were used to adjust the levels of all the words in order to normalize the hearing threshold as described in section 2.2.2.9.1. Because the wave files used in generating sentences had been normalized at an earlier stage (section 2.2.2.5.3) the total relative adjustments of the parts used to generate the sentences varied between -8.6 dB and +8.8 dB, but 90 % of the relative adjustments were in the range of -4 dB and +5 dB. Some of these adjustments are noticeable when listening to generated sentences, but this does not seem to reduce the quality substantially. A negative aspect of this procedure for normalization is that not all realisations of the same word will be adjusted to the same level. A better procedure for performing these normalizations will be proposed in section 6.7. The slopes of the performance-intensity curves for some of the words are rather shallow. Therefore, section 6.7 also presents a proposal for how this could be compensated for.

The performance-intensity curves for the word groups show that the verbs have a substantially poorer threshold than the other word groups. The slope of the performance-intensity curve for all the words has been found to be 10.6 %/dB (section 2.3.3.1.2). If we use the median value of the slopes for the individual words in equation (2.2), this value corresponds well with Kollmeier’s probabilistic model (presented in section 2.2.3.2.1).

No great variations have been found in terms of how well our speech material functions for the five main dialect groups of Norwegian (section 2.3.2.3). Other investigations support this view provided that a speaker from the Eastern dialect group is used, as we did when making the recordings for our material.

When producing the tests on CDs with one channel used for speech and the other for noise a temporal fixation of the noise and speech material is inevitable. In section 2.3.2.4 we described the indications that this fixation does not detract from the value of our measurements. Had we used modulated noise, however, we assume that this fixation might have been a problem.

In summary, the first field test confirmed that the five-word sentences generated by the diphone method had the qualities needed for further development of the new speech audiometry material.

2.4.3 Second field test

After the first field test, the words were normalized to the same threshold in noise. However, since there was an error in the normalization procedure deployed, the levels of the verbs and the adjectives used in the second field test were incorrect (see section 2.2.2.9.1). This error has been corrected at a later stage; thus, the levels are correctly adjusted for the words used in the subsequent tests and in the final material. To avoid difficulties associated with differences in the calibration of the audiometers, the levels in the second field test were measured as sensation levels. A reference sensation level for five-word sentences was established initially for each subject. The variations in speech recognition threshold without noise were measured. The standard deviation of the thresholds for the individual words was 2.0 dB, and the total variations in thresholds for the words were between -7.0 dB and +4.6 dB relative to the mean value. The mean value of the slope was found to be 10.9 %/dB. The values and the logistic-curves in Figures 2.29-2.33 show that the words also function well for speech recognition measurements without noise.

As shown in Figure 2.34, the thresholds and slopes for the word groups are very similar. It may seem surprising that the thresholds and slopes for the verbs and adjectives, which had incorrect levels, were approximately equal to the thresholds found for the other words, which were correctly adjusted. The reason may be that the level adjustments that were made were fairly small ones, and the basis for these adjustments were the measurements in the first field test with noise; whereas the measurements in the second field test were performed without noise.

Two test methods were compared: one with 5 dB descending steps, and the other with additional 1 dB rising steps. There was no great discrepancy between the methods, and the differences between the thresholds estimated with these two methods had a mean value of -0.19 dB, and a standard deviation of 1.12 dB (section 2.3.3.2).

Some measurements were performed to check whether the level adjustments performed after the first field test had altered the threshold in noise, but there was no evidence of any great differences for this adjusted material (section 2.3.3.3).

2.4.4 First laboratory test: threshold measured in hearing level

The speech recognition threshold had to be established in hearing level. The measurements were therefore performed in our department laboratory where the calibration of all the audiometers was checked prior to the measurements.

A binaural test was evaluated and a number of thresholds with standard deviation were collected for a small group of young normal-hearing subjects. These results were used to mark normal response areas in the protocol sheets used for this measurement.

2.4.5 Second laboratory test: threshold and slope in noise

The threshold and slope for five-word sentences in noise had not been established after the correct normalization of the levels of each word had been performed as described in section 2.2.2.9.1. The threshold was measured on young normal-hearing subjects and had a mean value of -6.0 dB SNR, with a standard deviation of 0.8 dB. The mean value of the slopes was 14 %/dB, with a standard deviation of 3.4 %/dB. The slope was more shallow than what has been found to be the case for the Swedish and German sentences, but a little steeper than for the Danish ones (Table 2.12). The reasons for these differences are not clear. The slope is very good compared to what was found in the first field test, where the median of the slopes for the individual words had the value 13.3 %/dB.

2.5 Conclusion

A Norwegian variant of Hagerman's sentences has been developed following a new diphone method which gives more natural-sounding speech than sentences produced by previous methods. The selected words are commonly understandable, with a phonemic balance close to the distribution of phonemes in the Norwegian language. The words were level-adjusted to achieve uniform speech recognition thresholds in noise. Measurement of speech recognition thresholds without noise demonstrated that the material was also well suited for speech audiometry measurements without noise, as there were no great differences between the thresholds for the individual words. The slopes of the performance intensity functions were steep both in silence and noise, but not as steep as for the Swedish and German sentences of the same type.

A Norwegian sentence test has been made available. Its performance on normal-hearing subjects has been documented.

Chapter 3

Three-word utterances

3.1 Introduction

For this new speech audiometry design for Norwegian we have decided to recommend using a new type of material for measuring the Speech recognition threshold (SRT). Traditionally, spondee words have been used for this purpose, both nationally and internationally. However, spondees are rare in Norwegian, and it is therefore difficult to select common words for inclusion in speech audiometry lists. During an informal discussion B. Hagerman made a proposal (personal communication, 2004): Why not use the three last words in the five-word sentences? Inspired by Hagerman's suggestion, this chapter presents speech audiometry material for SRT-measurements based on three-word utterances.

3.1.1 Spondee or not spondee

Traditionally, the speech recognition threshold has been measured by tests utilizing spondee words. A spondee is a word containing two syllables, both of which are stressed. Quist-Hanssen (1965) developed spondee lists for his speech audiometry. He states that spondees are relative rare in Norwegian: only a few are found among the 2000-3000 most frequent words and most of them occur only among the 5000-7000 most frequent words. The Norwegian spondee lists have been criticized amongst people working within audiology in Norway. The main criticism has been that many of the words in the lists seem very unfamiliar to today's Norwegians. This impression is confirmed by the data presented in Figure 3.1, which shows cumulative distributions for four types of Norwegian speech audiometry material.

As our basis for checking word frequencies we used text material made available on the internet by the University of Bergen (UiB) (2003 and 2006). This web site contains a list ranking the 10 000 most frequent words based on 150 million words in text. The web site also contains a complete list of 462 055 ranked words based on 14.6 million words of text. We used the UiB-material to check the ranking of words. First we checked the list of 10 000 words based on the 150 million words in text. We then searched the complete list based on 14.6 million words of text for the words not found in the first list.

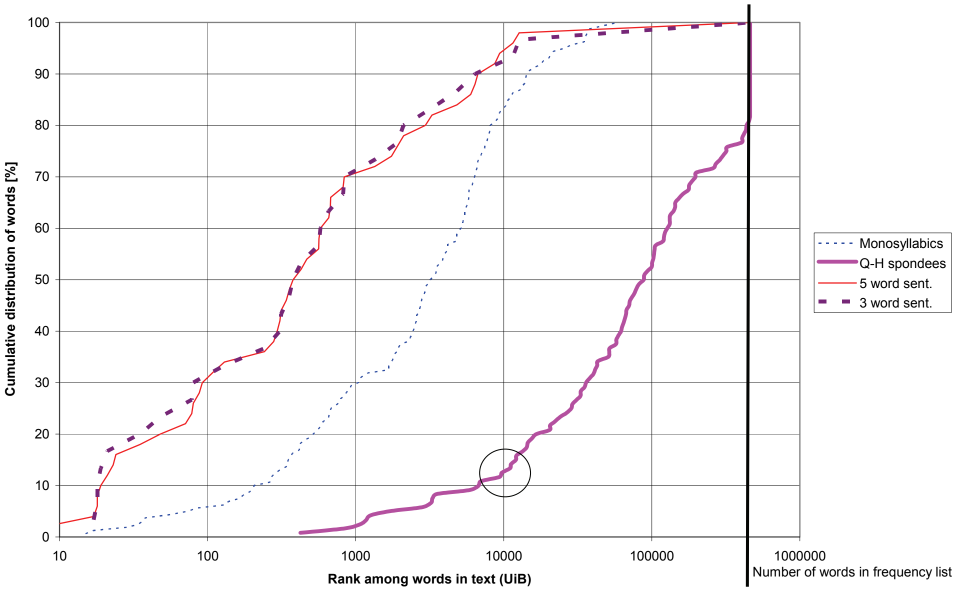


Figure 3.1 The cumulative distribution of monosyllabic words (Chapter 4), Quist-Hanssen (Q-H) spondee words, five-word sentences (Chapter 2) and three-word utterances as a function of their ranking among the UiB words.

The thick solid line in Figure 3.1 shows the cumulative distribution of Quist-Hanssen's 120 spondee words as a function of the ranking of UiB words. 23 of these words are so rare that they are not included in the UiB frequency list, which was limited to 462 055 words. This limit is indicated in the figure by the vertical line. Only 15 of the spondee words are among the 10 000 most frequent words (highlighted with a circle in the figure).

The thin dotted line in the middle shows the cumulative distribution of 160 monosyllabic words which will be presented in Chapter 4. The thin solid line shows the cumulative distribution of the five-word sentences based on 50 words presented in Chapter 2. The thick dotted line shows the cumulative distribution for three-word utterances based on the 30 words needed to construct these sentences.

Figure 3.1 shows clearly that both the selected monosyllabic words and the words used to make the five-word sentences are much more frequent than Quist-Hanssen's spondee words. Even if we changed the majority of the spondee words the cumulative distribution would still be unsatisfactory if our goal is to have a list of frequent words. And it is indeed our goal to use frequent words, since we want to measure speech recognition threshold – not the listeners' knowledge of rare words.

Since the frequency of spondee words is rather limited in Norwegian we decided to propose using this new type of speech material for SRT-measurements. A bonus associated with our method is increased measurement reliability due to the fact that we use three-word sentences, each of which counts as three items when registering the score, instead of a spondee word which counts as only one item, while the measurement time is approximately the same.

3.2 Methods

3.2.1 Speech material

The decision to use the last three words of the five-word sentences meant that the already existing material as described in Chapter 2 could be reused as the basis for this new material. Thus all the material was almost ready for use, requiring only some finalizing before making test CDs to be used to evaluate measurements of speech recognition thresholds with three-word utterances.

For each of the words in the three-word utterances we have 10 alternatives. The number of possible utterances is computed as:

10 numerals · 10 adjectives · 10 nouns = 1000 utterances

Since each list will be composed by using each of the 30 selected words only once, these 1000 utterances can generate 100 lists of 10 utterances each with no identical utterances.

3.2.1.1 Utterances made by using the last three words of the diphone material

The words used for generating the new speech material are presented in Table 3.1 (cf. Chapter 2.2.1.4).

Table 3.1 The Norwegian words selected for generating three-word utterances (*English translation*).

numeral	**adjective**	**noun**
to (*two*)	gamle (*old*)	knapper (*buttons*)
tre (*three*)	hele (*whole*)	boller (*muffins*)
fire (*four*)	store (*big*)	vanter (*gloves*)
fem (*five*)	nye (*new*)	penner (*pens*)
seks (*six*)	vakre (*pretty*)	kurver (*baskets*)
sju (*seven*)	mørke (*dark*)	skåler (*plates*)
åtte (*eight*)	lyse (*bright*)	luer (*caps*)
elleve (*eleven*)	fine (*fine*)	duker (*tablecloths*)
tolv (*twelve*)	lette (*light*)	ringer (*rings*)
atten (*eighteen*)	svarte (*black*)	kasser (*boxes*)

3.2.1.2 Phonemic balance

Figure 3.2 shows the distribution of the phonemes of the 30 words from Table 3.1 used to generate the three-word utterances (solid line with diamonds). For comparison the distribution of the five-word sentences from Chapter 2 is also shown (dotted line with triangles). Finally the distribution of the phonemes from the Bergen material based on the 20 000 most frequently used words is included (columns) for reference. The procedure for computing this distribution is described in section 2.2.1.3.

As is evident in Figure 3.2 there is a greater difference in the phonemic balance between the three-word utterances and the distribution found in the UiB material than between the five-word sentences and this material. Only one phoneme (A) is closer to the UiB material in this respect; about 14 phonemes (d, k, j, m, n, l, r, i, @, e:, A:, y:, }: and {i) are slightly further away from the UiB material; and the remaining 36 phonemes have approximately similar distributions in the three- and five-word sentences. A lower score for phonemic balance for the three-word utterances is as expected since the they consist of only 30 words, and these 30 are a subset of the 50 words used for the five-word sentences.

A better phonemic balance score for the 30 words used to generate three-word utterances could have been achieved by substituting some of the words selected with other words, but this was not an issue since the 30 words are a subset of the 50 words used in Chapter 2 and the decision to use three-word sentences was made after the five-word sentences had been realized and evaluated.

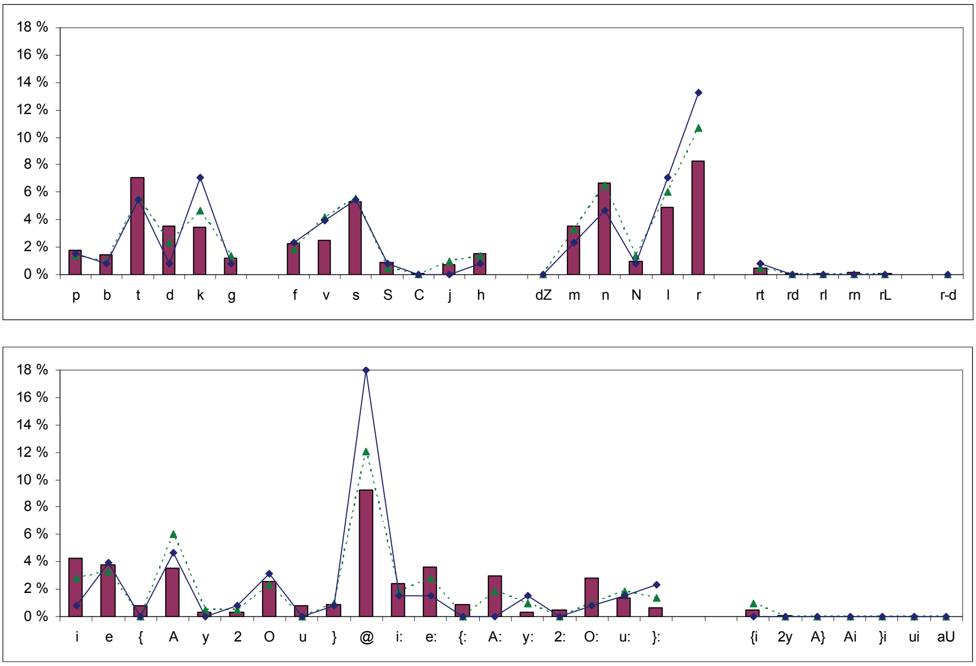


Figure 3.2 Distribution of phonemes: Three-word utterances (solid line with diamonds), the five-word sentences from Chapter 2 (dotted line with triangles) and the 20 000 most frequent words in the University of Bergen material, corrected for the frequency of each word (columns).

3.2.2 Preparation of the stimulus material

Table 3.2 The wave files needed to generate five-word sentences and three-word utterances.

Word type	Name	verb		numeral		adjective		noun
Example sentence	*Thea*	*lå-*	*-åner*	*se-*	*-eks*	*va-*	*-akre*	*kurver*
Five-word sentence wave files	wave file type 1		wave file type 2		wave file type 3		wave file type 4	
Three-word utterance wave files				wave file three-word utterance start			wave file type 4	

From the five-word sentences developed using the diphone method as described in Chapter 2 we took the last three words and used them to develop this new three-utterance material. Table 3.2 clarifies which wave files were needed to generate these three-word utterances and five-word sentences. We had 100 recordings of each of the four types of wave files needed to generate the five-word sentences (Table 3.2, third row). In order to generate the three-word utterances we could keep wave file type 4 unchanged, but we needed a new type of wave files for the start of the three-word utterances (Table 3.2, bottom row). To make this new type of wave files we had to combine wave files type 2 with the matching wave files type 3. The treatment and analysis of the wave files described in Chapter 2 had provided us with information about the timing of the transitions between the words for these wave files. We used a Matlab routine to fetch the wave files needed to generate the five-word sentences. A wave file type 2 containing the last part of the second word (verb) and the first part of third word (numeral) was fetched together with a wave file type 3 from the same recorded sentence containing the last part of the third word (same numeral) and the first part of the fourth word (adjective). These wave files were concatenated, and the first part containing the last part of the second word (verb) was stripped from the file before the resulting wave file was saved under a new name. This procedure was repeated until we had the one hundred wave files we needed, containing all the combinations of the third word (numeral) with all the combinations of the first part of the fourth word (adjective). Each of these one hundred wave files could then be combined with ten of the one hundred wave files type 4 from the five-word sentence material containing the last part of fourth word (same adjective) and the fifth word (noun).

3.2.3 Listening tests

Three listening test sessions were performed with the three-word utterances. In the second field test, thresholds and slopes for words without background noise were measured in sensation level (dB SL). The design of the second field test is presented in the following section. In order to establish thresholds and slopes for words without background noise in hearing level (dB HL) we conducted a listening test in the laboratory of the Audiology Programme. The design of this test, the first laboratory test is presented in section 3.2.3.2. Finally the second laboratory test presented in section 3.2.3.3 was designed to get information about the threshold and slope for three-word utterances in noise.

3.2.3.1 Second field test: threshold and slope for words without noise, measured in sensation level

This listening test was performed in the second field test described in section 2.2.3.3. The actual test was conducted by students from the

Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, during their practice period in May-June 2006. The second field test incorporated test items from the five-word sentences as described in Chapter 2, plus test items from the three-word utterances described in this chapter, as well as test items from the monosyllabic words described in Chapter 4.

The main objective of using three-word utterances was to check the hearing thresholds for the individual words without background noise. The levels of the words were adjusted based on the results from the first field test performed using five-word sentences with noise.

As described in section 2.2.3.3, it was not possible to establish correct calibration in this field test. Therefore the results cannot be expressed in hearing level (dB HL), but will instead be presented in terms of sensation level (dB SL), for which the reference threshold for each subject was established for the five-word sentences.

An example of the scoring protocol can be found in Appendix C. Seven test sets (A, B, C, D, E, F and G) were made with equal structure but different sentences and words.

The students were instructed to use the scoring protocol in Appendix C and tracks 20 to 34, which contained 10-sentence lists of the three-word structure without background noise. The measurement procedure was to start at such a high level that more than 80 % of the words were identified for the first two lists. This guaranteed that the test persons became somewhat familiar with the words. The level was reduced by 5dB after each list until the score dropped below 20 %. Then the level was to be raised by 1 dB steps until the score again exceeded 80 % and this part of the test was finished. 34 test persons with normal hearing participated in this test.

The results were evaluated by first determining the individual threshold for each subject for the five-word sentences as described in section 2.2.3.3.1. Then all the measurements were pooled together in 1 dB bins so that the responses for each word at different sensation levels could be used to estimate the logistic function for each word. The logistic function was estimated with the method of least squares by using the solver which is a standard add-in in Excel.

3.2.3.2 First laboratory test: threshold and slope for words without noise, measured in hearing level

This test was performed by students from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, in the department's laboratory during September 2007. The laboratory test incorporated test items from Quist-Hanssen's speech audiometry, three-word utterances, monosyllabic words, five-word sentences and three-numerals lists.

The main objective of the testing of three-word utterances was to check the hearing thresholds for the individual words without background noise. Since this was an in-house test, all the speech audiometers were calibrated before the testing and the results can thus be presented in terms of hearing level (dB HL) unlike the results from the preceding section which were expressed in terms of sensation level (dB SL).

An example of the scoring protocol can be found in Appendix D. Five test sets (A, B, C, D and E) were made, with equal structure but different sentences and words.

The students were instructed to use the scoring protocol in Appendix D, which included two tests using three-word utterances (cf. page 2 and page 4 in protocol). Page 2 was used for testing tracks 5 and 6, which comprised quick-speed tests using three-word utterances. Each of the tracks was a list of 30 utterances, where each new utterance was reduced by 1.5 dB relative to the one preceding it. The pure-tone-average hearing level in the table on page 3 of the scoring protocol provided the starting level for these tests. Page 4 of the protocol was for testing tracks 7-16, where each track was a three-word utterance list of 10 utterances. These tracks were meant to measure points on the performance-intensity function between 0-100 % score with a suitable level selected for each track. 19 young subjects with normal hearing participated in this test.

The results were pooled together in 1 dB bins so that the responses for each word at different sensation levels could be used to estimate the logistic function for each word. For each word we had between 238 and 290 tests. The logistic function was estimated with the method of least squares by using the solver which is a standard add-in in Excel. The squared errors where weighted with the number of measurements in each bin for this calculation.

For estimation of the speech recognition threshold in order to determine the calibration level, the three-word utterances on page 4 of the protocol were used. The results were pooled together and a logistic function was estimated in order to determine the speech recognition threshold. These measurements could be compared with similar measurements for monosyllabic words on page 5 and digit triplets on page 9 of the protocol.

For this test a calibration signal 1 dB lower than the equivalent levels measured without the weighting filter of the speech material had been calibrated to 20 dB SPL in the earphones of the audiometer. The calibration level of the final material was altered after this test as described in section 2.2.3.6.

3.2.3.3 Second laboratory test: threshold and slope in noise

This test was performed by the author on students and staff members from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, in the laboratory of the department during April 2008. The laboratory test incorporated five-word sentences, three-word utterances and monosyllabic words. The measurements were performed on one ear for each test subject, on a clinical audiometer calibrated for HiST speech audiometry. Noise from the speech audiometry CD was mixed ipsilaterally with the speech signals. Nine subjects participated in the test.

The three-word utterances were presented in lists containing 10 sentences each. After having the subjects listen to one list without noise, the remaining lists were presented at the signal-to-noise ratios of 10, 0, -4, -8 and -12 dB SNR, and the number of correctly recognized words was registered for each utterance. For each subject a different list was selected as the first one and then the remaining list followed in the same order as the tracks on the CD.

A sigmoid function was fitted to each subject's response by the least squares method using the solution solver in Excel. The mean values and standard deviations of the slope and threshold are presented in 3.3.3.

3.2.4 Preparation of the final material

As described in section 2.2.2.9.1 an error was revealed in the level adjustments of the adjectives used in the second field test were revealed after the completion of the test. Correct adjustments according to the intentions described in section 2.2.2.9.1 have later been made so that all the words used in the subsequent tests and in the final material have been subjected to the correct level adjustment.

For the generation of the three-word utterances 200 wave files are available. Realization of the 100 lists containing all of the available 1000 utterances according to this method is therefore possible. A nomenclature has been developed to control the lists and the sentences realized in order to avoid possible repetitions of utterances. The nomenclature is presented in Appendix F.

Matlab routines were developed for the generation of new wave files, usually containing a complete list of ten utterances in one channel and optional noise sequences in the other. A list was selected from the nomenclature, and the order of the utterances was randomized in a spreadsheet containing a text table of and codes for the utterances. The text table was used in the documentation of the tests. The codes for the utterances were imported into the Matlab routine. The Matlab routine then automatically selected and concatenated the wave files necessary for generating each utterance, adjusted the levels for each utterance in accordance with the test procedure, inserted pauses between the utterances, and optionally inserted noise sequences in the other channel at the correct positions. The noise was fetched from the noise wave file described in section 2.2.2.6 with a random start position within that file.

3.3 Results

3.3.1 Second field test: threshold and slope for words without noise, measured in sensation level

The results from the listening tests were evaluated using Excel following the procedures outlined in 3.2.3.1.

3.3.1.1 Individual words

Figures 3.3-3.5 on the following pages show the fitted logistic function for each word as a function of the sensation level. The sensation level was referenced to the speech recognition threshold for five-word sentences for each subject. The mean SRT value across all words is -0.9 dB SL, with a standard deviation of 2.3 dB. The largest differences from the mean are the values achieved for the noun *boller*, which has an SRT 4.4 dB better than the mean, and for the noun *kasser*, which has an SRT 4.9 dB worse than the mean (Figure 3.5).

The mean and median slopes of the words are very similar at, respectively, 11.4 %/dB and 11.7 %/dB. The standard deviation is 2.8 %/dB. The variations are within 6.1 %/dB for the noun *luer* (Figure 3.5) and 19.6 %/dB for the numeral *to* (Figure 3.3).

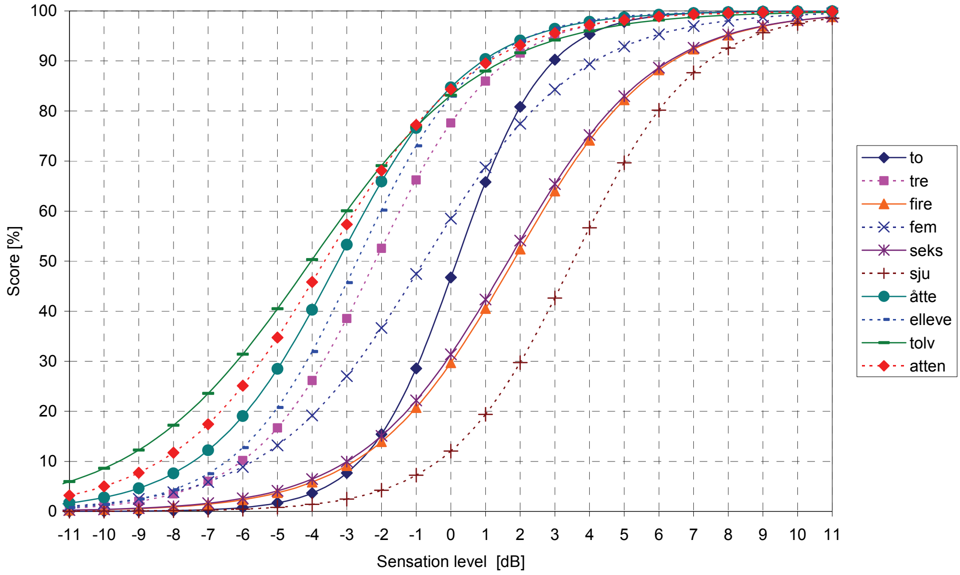


Figure 3.3 The fitted logistic function for the numerals as a function of sensation level.

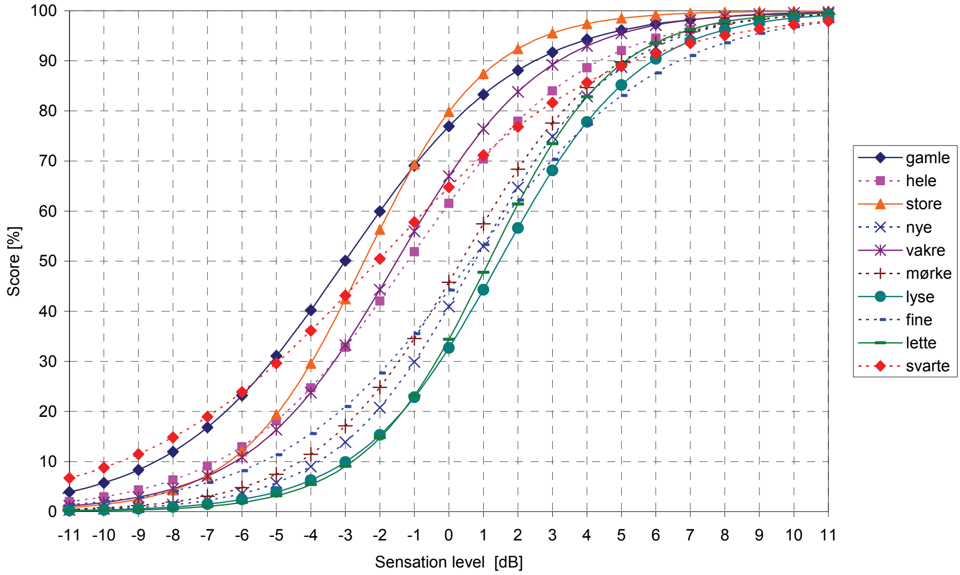


Figure 3.4 The fitted logistic function for the adjectives as a function of sensation level.

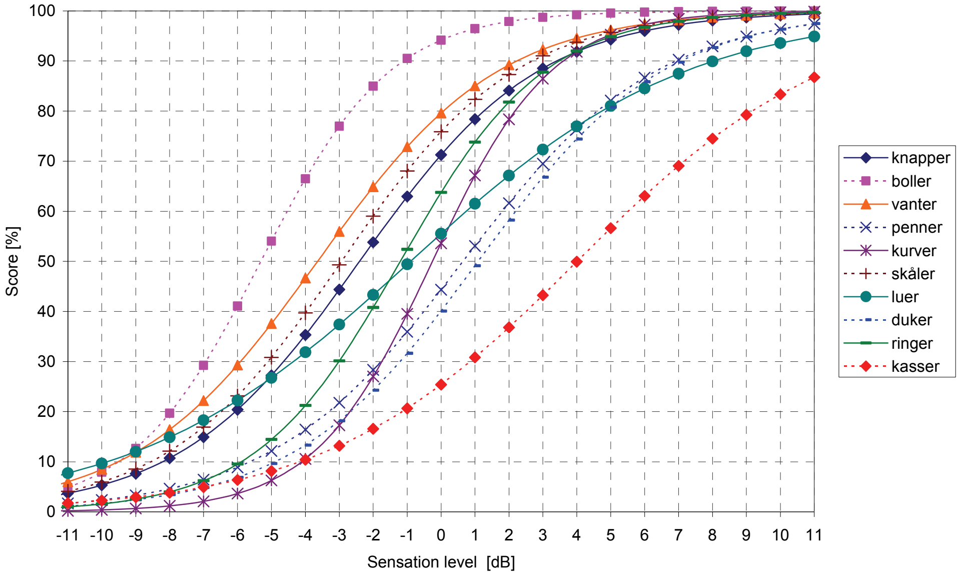


Figure 3.5 The fitted logistic function for the nouns as a function of sensation level.

3.3.1.2 Word Groups

The fitted logistic curves for numerals, adjectives and nouns are very similar and lie close to the fitted logistic curve for all the three-word utterance words (Figure 3.6).

The SRT scores obtained using numerals, adjectives and nouns differ by, respectively, -0.1, -0.4 and -0.3 dB from the SRT score for all words. The slopes of numerals, adjectives and nouns are, respectively, 9.6, 10.0, and 8.2 %/dB; and for all the words the slopes of the fitted logistic function is 9.2 %/dB.

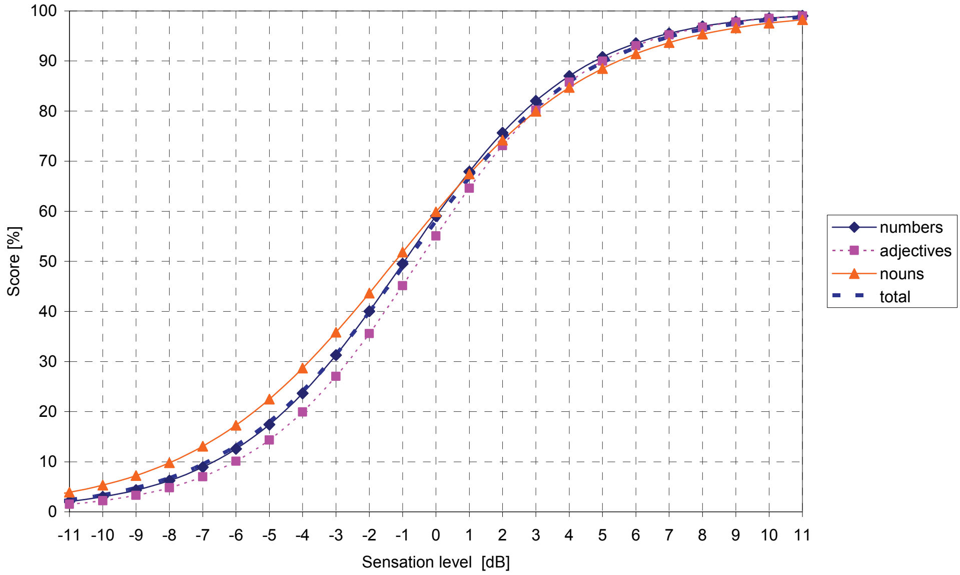


Figure 3.6 The fitted logistic function for the word groups and for all the words as a function of sensation level.

3.3.2 First laboratory test: threshold and slope for words without noise, measured in hearing level

The results of the listening tests were treated in Excel following the procedures outlined in 3.2.3.2

3.3.2.1 Individual words

Figures 3.7-3.9 on the following pages show the fitted logistic function for each word as a function of the hearing level. The mean SRT value across all words is 1.9 dB HL with a standard deviation of 1.8 dB. The largest differences from the mean are the values achieved for the noun *boller,* which has an SRT 3.2 dB better than the mean (Figure 3.9), and the numeral *sju,* which has an SRT 3.8 dB worse than the mean (Figure 3.7).

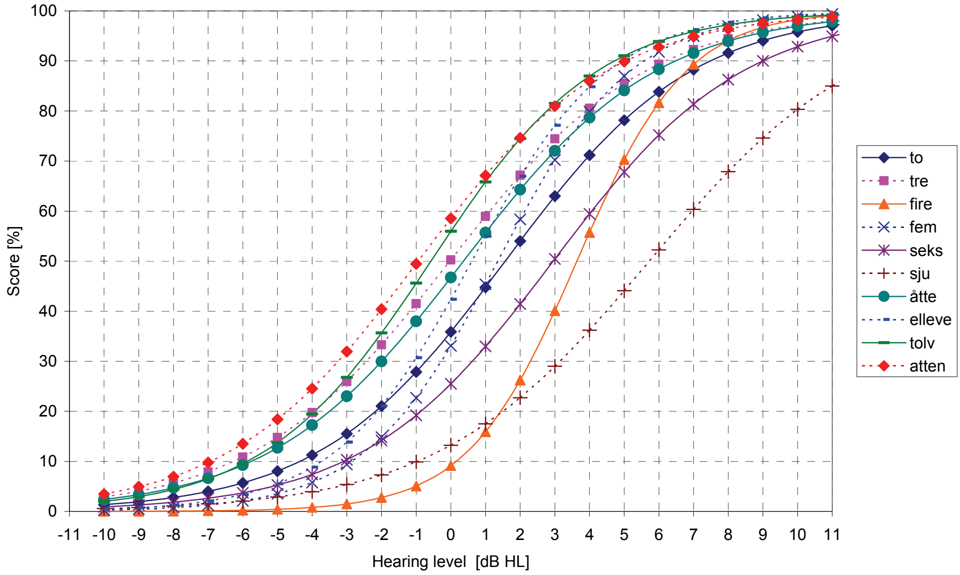


Figure 3.7 The fitted logistic function for the numerals as a function of hearing level.

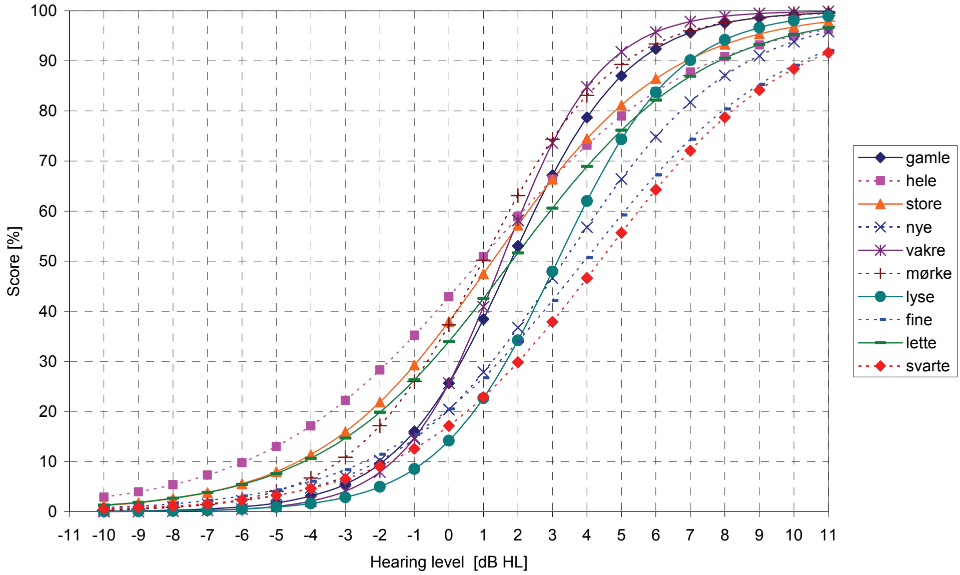


Figure 3.8 The fitted logistic function for the adjectives as a function of hearing level.

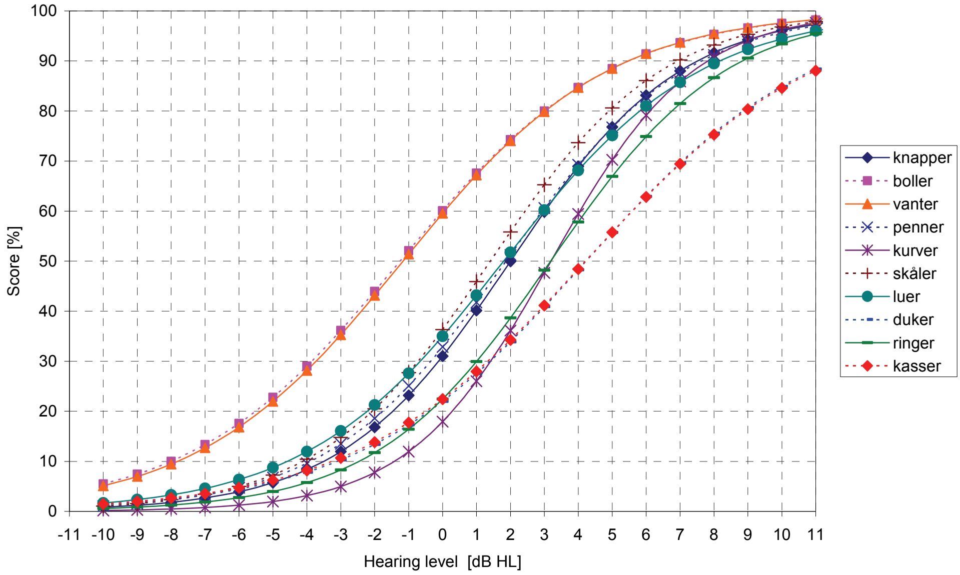


Figure 3.9 The fitted logistic function for the nouns as a function of hearing level.

The mean and median slopes are respectively 10.4 %/dB and 9.4 %/dB. The standard deviation is 2.6 %/dB. The variations are within 7.4 %/dB for the noun *kasser* (Figure 3.9) and 17.4 %/dB for the adjective *vakre* (Figure 3.8)

3.3.2.2 Word Groups

The fitted logistic curves for numerals, adjectives and nouns are very similar and lie close to the fitted logistic curve for all the three-word utterances (Figure 3.10).

The SRT scores obtained using numerals, adjectives and nouns differ by, respectively, -0.5, 0.3 and 0 dB from the SRT score for all words. The slopes of numerals, adjectives and nouns are, respectively, 9.0, 10.3, and 8.5 %/dB; and for all the words the slopes of the fitted logistic function is 9.3 %/dB.

3.3.2.3 Speech recognition level for calibration

For the calibration of all the material in “HiST taleaudiometri” (HiST speech audiometry) a common procedure was developed – one that could be used for the three-word utterances, the monosyllabic words and the digit triplets. The threshold was estimated by fitting a logistic curve to the selected measurements, and the threshold was found to be 2.3 dB HL. There is a small difference (0.4 dB) between this threshold and that found across all words in section 3.3.2.1. The reason for this discrepancy is that here only the 10-utterance lists on page 4 of the score protocol (Appendix D) was used, whereas in section 3.3.2.1 the quick-speed lists on page 2 of the score protocol were used in addition. The calibration level has later been changed as reported in section 2.2.3.6.

The speech recognition level was found to be 1.0 dB better for three-word utterances than for five-word sentences when the quick-speed tests were applied to both sets of material.

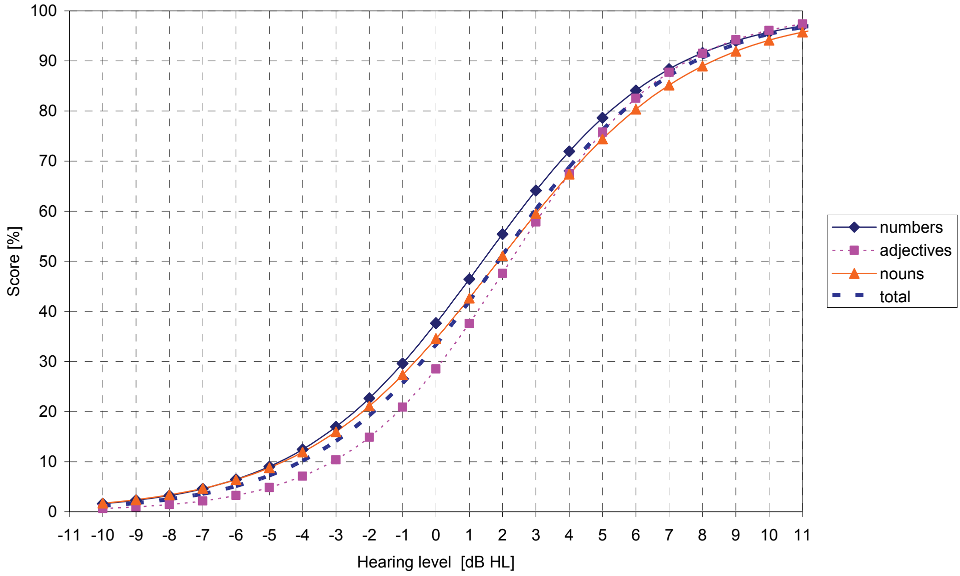


Figure 3.10 The fitted logistic function for the word groups and all the words as a function of hearing level.

3.3.3 Second laboratory test: threshold and slope in noise

Measurements and analysis of the results were carried out as described in section 3.2.3.3. The mean value of the threshold was -6.2 dB SNR, with a standard deviation of 0.8 dB. The mean value of the slope was 16 %/dB, with a standard deviation of 3.5 %/dB.

The standard deviation of the slope is rather large. In order to check how much of this variability is inherent in the measuring method it was decided to do a simulation of the measurements according to the methods to be described in Chapter 5. A threshold of 35 dB and a slope of 16 %/dB were selected for the simulation. The same levels relative to threshold as given in section 3.2.3.3 were used with ±1 dB randomization. Instead of using 30 words in the test list we reduced this number to 28 due to the context effect found for these lists (to be described in section 5.2). The simulation method is described in section 5.3.6 and gives the expected mean value and standard deviation for both threshold and slope. The results of 500 simulations are shown in Figure 3.11.

The standard deviation of the slope for the simulations of the hypothetical subject was 3.3 %/dB. In our laboratory measurements the standard deviation was 3.5 %/dB. These values are almost equal - which is an indication that most of the variability in the slope is related to the measurement method and consequently only to a small degree to individual differences between the subjects.

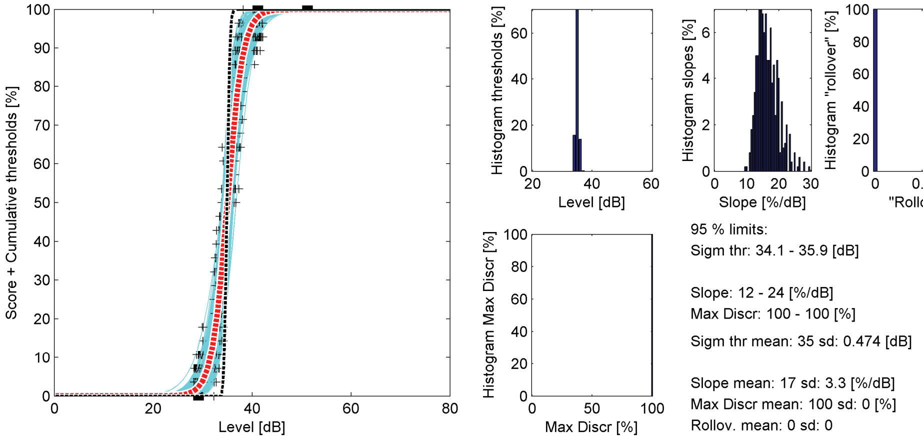


Figure 3.11 Hypothetical subject with slope 16 %/dB simulated measured by fitting a logistic curve to the scores. 28 test items in each set measured at 5 levels. The large panel shows the logistic function for the hypothetical subject, indicated by the thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the fitted logistic curves of the scores. The medium dashed line shows the cumulative distribution of the threshold estimated by the curve fitting routine. The small top left panel shows the histogram of the thresholds obtained during the 500 simulations by the curve fitting routine. The small top middle panel shows the histogram of the estimated slopes. The small top right panel shows the histogram of the estimated rollover parameter. The small bottom panel shows the histogram of the estimated maximum recognition score. The 95 % limits and/or means plus standard deviations of the estimated parameters are indicated.

3.4 Discussion

3.4.1 Second field test: threshold and slope for words without noise, measured in sensation level

After the first field test, the noise threshold was normalized to the same level for all the words, except for the fact that there was an error in the deployed normalization procedure, resulting in incorrect levels for the adjectives used in the second field test (section 2.2.2.9.1). This error was corrected at a later stage so that the words used in the subsequent tests and in the final material have correct levels. For the five-word sentences a reference sensation level was initially established for each subject. The mean SRT value for all the words is 0.9 dB SL, with a standard deviation of 2.0 dB. The threshold variations for the words were in the range of -4.4 dB to +4.9 dB relative to the mean value. The mean value of the slope was found to be 11.4 %/dB. The values and the logistic curves in Figures 3.3-3.5 show that the results for all the words are rather similar, without large differences in threshold and slope. If we compare with Figures 2.31-2.33 showing the same curves for the words used in the five-word sentences, we see that the behaviour is almost identical. These results are promising and indicate that the three-word utterances offer a very plausible alternative basis for speech recognition testing.

The thresholds and slopes for the word groups are very similar, as shown in Figure 3.6.

3.4.2 First laboratory test: threshold and slope for words without noise, measured in hearing level

This test was performed using the correctly normalized words and the variation in the thresholds for the different words was even better than what was found in the second field test. All the thresholds for the individual words were located within the area of between -3.2 dB to +3.8 dB from the mean value. This represents a range that is more than 1 dB tighter in both ends compared to the variation between -4.4 dB and +4.9 dB recorded in the second field test measured in sensation level. The mean slope was 10.4 %/dB. By comparing Figures 3.7-3.9 with 3.3-3.5 it is evident that all the word groups – numerals, adjectives and nouns – are more tightly matched when measured in hearing level than when measured in sensation level. One reason for the more tightly match in the first laboratory test could be that all the words were now correctly normalized, contrary to the second field test where the adjectives had received incorrect level adjustments. However it does not seem reasonable that this difference should give a tighter range also for the numbers and nouns. The most reasonable explanation could be that for the second field test measurements in sensation level, we had to establish an individual threshold reference level for five-word sentences which where used to indirectly give the results of other measurements on the same subject in dB SL. Whereas for the first laboratory test we had control over the audiometer calibration so all the measurements could be directly expressed in dB HL. It can expected that the measurements performed in the second field test which required two steps are not as accurate as the direct approach used in the first laboratory test.

Figure 3.10 shows that the performance-intensity curves for the three word groups match very closely.

The measurements confirm that three-word utterances offer a very plausible alternative basis for speech recognition threshold measurements.

3.4.3 Second laboratory test: threshold and slope in noise

The second laboratory test measured the threshold and slope for three-word utterances in noise. The threshold was measured on young subjects with normal hearing and had a mean value of -6.2 dB SNR, with a standard deviation of 0.8 dB. The mean value of the slopes was 16 %/dB, with a standard deviation of 3.5 %/dB. Much of the variability is a result of the low accuracy of the measurement method used to estimate the slope. The slope is steeper than the slope of 14 %/dB found for the five-word sentences.

This test shows that the three-word utterances also represent a very plausible alternative basis for measurements in noise.

3.5 Conclusion

This chapter has described the development of three-word utterances, which represent a new type of speech audiometry material for speech recognition threshold measurements. The structure of the utterance is numeral-adjective-noun, and the words used are the same as the last three words of the five-word-sentences described in Chapter 2. The thresholds without noise for the individual words were good and without large deviations. The slopes of the performance-intensity function are steep both in silence and with noise.

The material is documented on people with normal hearing and was found to represent a plausible alternative basis for measuring speech recognition threshold both in silence and with noise.

Chapter 4

Monosyllabic words

4.1 Introduction

A new set of monosyllabic words has been developed as part of the “HiST taleaudiometri” (HiST speech audiometry). As described in section 1.2 the monosyllabic words in Quist-Hanssen’s speech audiometry have had a central position in Norwegian speech audiometry practice. When developing a new speech audiometry the decision was made not to continue the traditional practice at some institutions of conducting threshold measurements using the monosyllabic words. The new set of monosyllabic words was selected for the purpose of making speech intelligibility measurements as accurate as possible.

4.2 Methods

4.2.1 Speech material

Margolis and Millin (1971) summarize the traditional criteria used when developing material for articulation testing as proposed by both Egan and Hirsh, who designed such material in 1948 and 1952, respectively (section 1.1): First, use monosyllabic words. This reduces the number of clues available to the listener and ensures that the response is mainly dependent on the discrimination capacity. Second, use familiar words so that the listener’s level of education will not influence the test. Third, make the content of the lists representative of the syllable types and phonemes of the language, i.e. phonetically balanced (PB). Finally, make the tests equal in range and average level of difficulty so that the lists are equivalent. For the most part there is consensus about the first two as well as the fourth criterion, but the issue of phonetic balance is questioned by several researchers.

Lyregaard (1997) discusses the terms related to this question and states that the use of the term *phonetic balance* is incorrect: The appropriate term is *phonemic balance*, according to Lyregaard, because the goal is to achieve test material with the same phonemic balance as in everyday speech. Lyregaard also discusses the relationship between the lists and the test material as a whole, and proposes to use the term *phonemic equalization* if each of the lists of the test has the same phonemic balance, meaning that they can be considered as interchangeable. Tobias (1964) states that there is *"overwhelming clinical and experimental evidence that indicates phonetic balance to be an interesting but unnecessary component of one of our current audiometric tests"*. Carhart (1965) declares that, although phonetic balance may seem unnecessary, a broad distribution of phonetic content appears to be a requisite for a clinically valid discrimination test. Jerger (1970) points to the lack of progress in speech audiometry over the past 20 years (1950-1970) and says the following during the discussion at a Danavox symposium on the topic of speech audiometry:

> *"Traditional speech audiometry still has only limited diagnostic value, still can not distinguish among hearing aids, and still cannot tell us how much difficulty the patient faces in understanding real speech in the real environment. I suggest that this lack of progress is due to the fact the traditional approach to speech audiometry is based on the false assumption, namely that the critical listening ability for understanding speech is frequency or spectral discrimination. All of our traditional materials - nonsense syllables, monosyllabic phonetically balanced words, rhyming consonants - all are based on this oversimplified assumption that distinguishing among phonemes with similar acoustic spectra is essential to speech understanding. I think that two lines of evidence converse to suggest that this approach has led us up a blind alley. First, as Mr. Juhl Pedersen demonstrated yesterday, one may distort the frequency spectrum of speech far beyond that encountered in patients with hearing loss and yet the ability to understand running speech remains remarkably good. This is an important fact that we ought to think about very hard. Second, in spite of years and years of experimental efforts this traditional approach has brought us no closer to our long range goal, which is describing and measuring true communication handicaps. Our traditional techniques give us only the most gross approximation in this area. I suggest that we shall not make significant progress in the further refinement of speech audiometry, until we abandon the concept that speech audiometric material must have phonemic discrimination ability. On the contrary it is becoming increasingly clear that the key parameter for speech intelligibility is time. Temporal, not spectral,*

> *characteristics carry the information important for the understanding of real speech." (Jerger 1970, p. 233)*

Martin et al. (2000) compare results of speech audiometry performed with PB lists and with lists made up by randomly selecting words from a dictionary. They find no clinically meaningful differences between the lists. This is in contrast with Lyregaard et al. (1976), who pointed out that the phonemic equivalence of sublists is of crucial importance.

Jusczyk and Luce (2002) present a review of the past half century of research on speech perception and discuss the perceptual units of speech. Early studies focused on the phonetic segment as a minimal sound unit of speech, and researchers assumed that there were direct acoustic correlates between such units. When researchers sought to find the acoustic features corresponding to the phonetic segments, they discovered that it was impossible to divide the formants of a CV syllable into pieces corresponding to each segment. The phonetic segments were coarticulated. Later studies have brought no strong consensus regarding the basic perceptual unit. A range of units, such as demisyllables, context-sensitive allophones, syllables and context-sensitive spectra each has its supporters.

Both Quist-Hanssen(1965) and Slethei (1975) paid much attention to the phonemic balance of the lists they produced when developing their Norwegian speech audiometry. The references in the paragraphs above, however, do not support the need for phonemic balance. For the designing of the new set of monosyllabic words for "HiST taleaudiometri" the following strategy was chosen: First, to record the lists available from Quist-Hanssen and three lists of words for children used by Rikshospitalet University Hospital in Oslo (sections 4.2.2-4.2.3). Second, to test which of the words are more easy or more difficult to recognize in a simple listening test (section 4.2.4.1). Third, to evaluate the words for inclusion based on both the listening test and the frequency of the words (section 4.2.5.1). Finally, to mix the words in 50-word lists with the best possible phonemical equalization and distribution of easily and less easily recognizable words between the lists (section 4.2.5.3). Since Quist-Hanssen's monosyllabic words, which represent the main source of words, were phonemically balanced, the new lists will also have a good phonemic balance (section 4.2.5.2).

4.2.2 Sources of words

4.2.2.1 Quist-Hanssen words

The speech audiometry material developed by Sverre Quist-Hanssen (1965) for Norwegian in the 1950s includes 170 monosyllabic words. According to Quist-Hanssen, the words were selected among the first 3000-6000 words used by children in primary school, and most of the selected words were among the first 2000, and thus well-known to everybody.

There is no available material for checking word frequencies in child language, so we had to depend on more general material. We decided to use the word frequency material from the University of Bergen (UiB) (2003 and 2006) presented in section 2.2.1.2.

Figure 4.1 shows the cumulative distribution of Quist-Hanssen's monosyllabic words as a function of the ranking of UiB words. In this figure we see that about 90 % of the Quist-Hanssen words are found among the 20 000 most frequent words, and that half of his words are among the 3500 most frequently used words (highlighted by the two circles in the figure). The least common word among the ones selected by Quist-Hanssen is ranked as number 418 081. This fact contradicts Quist-Hanssen's statement, referred to in the first paragraph of this section, that his words were chosen among the first 3000-6000 words used by children in primary school. The discrepancy between Quist-Hanssen's statement and the UiB-material may be partially due to differences in the frequency of words between text and children's language, and to language development over time.

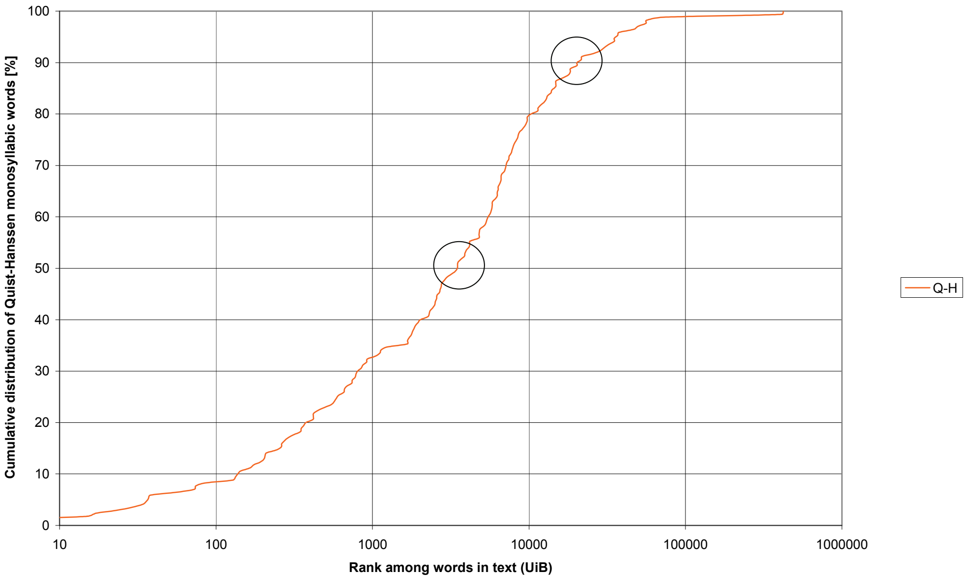


Figure 4.1 The cumulative distribution of Quist-Hanssen (Q-H) monosyllabic words as a function of their ranking among the UiB words.

4.2.2.2 Words for children selected by Rikshospitalet

Some of Quist-Hansen's words are deemed too complicated for performing speech audiometry with monosyllabic words on children. Rikshospitalet University Hospital in Oslo has therefore selected new monosyllabic words for this usage. The material was produced by their audiologists and speech-language pathologists, who are experienced in the audiological testing of children. In cooperation with Rikshospitalet University Hospital we have made recordings of this material. There are 3 lists, each consisting of 50 monosyllabic words. Children's list 1 (RC1) contains words appropriate for small children; children's list 2 (RC2) contains the same words as list 1 but in a different order. Children's list 3 (RC3) is for children who are somewhat older. It contains some of the words from list 1, supplemented with many words from the Quist-Hanssen list, plus a few other words.

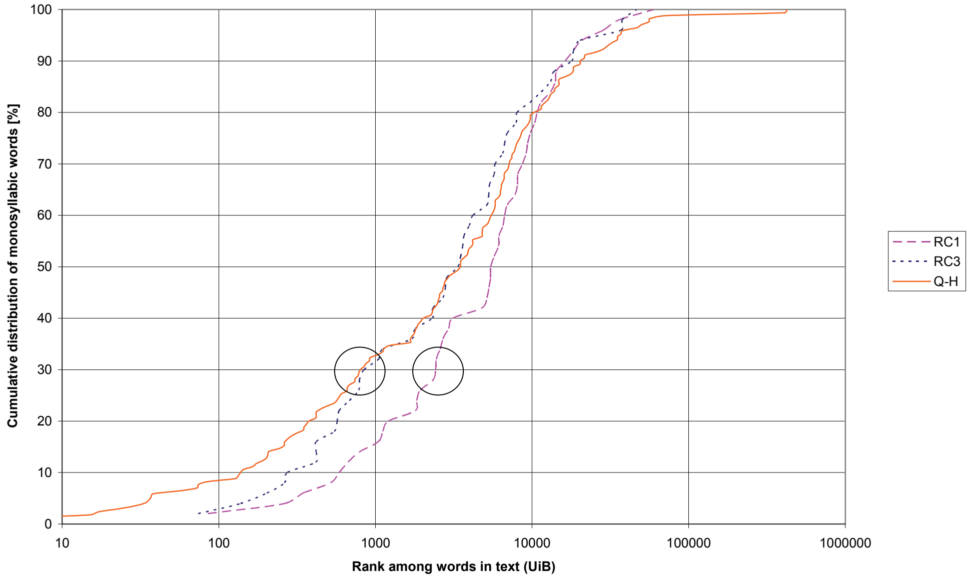


Figure 4.2 The cumulative distribution of Rikshospitalet's monosyllabic lists for children (RC1 and RC3) and Quist-Hanssen's (Q-H) monosyllabic words as a function of their ranking among the UiB words.

A comparison of the selection of words for RC1 and RC3 with the selections made by Quist-Hanssen is shown in Figure 4.2. Here we use the same technique as in Figure 4.1 in order to compare the cumulative distribution of RC1 and RC3 with the selections made by Quist-Hanssen. In Figure 4.2 we notice that the cumulative distribution of RC1 shows the use of words that are less frequent than both those found on RC3 and among the Quist-Hanssen words. For example, 30 % of the RC1 words are among the

2500 most frequently used words. In comparison, 30 % of the RC3 words or the Quist-Hanssen words are taken from the 800 most frequently used words (highlighted by the two circles in the figure). This demonstrates that the UiB-material fails to provide a good representation of the word frequencies of children's language – which is understandable since the RC1 material was specially developed for small children.

But even if the rise of the distribution curve for RC1 is delayed compared to Quist-Hanssen's words it increases more steeply, so that the least common words near 100% in the figure are more highly ranked for the child language lists than for the Quist-Hanssen material. The least common word in RC1 is ranked as number 59 193. In comparison, the least common of Quist-Hanssen's words is ranked as number 418 081.

Based on the results examined in the preceding paragraphs we can conclude that the cumulative distribution of words ranked according to the UiB material can be of help in selecting common words for speech audiometry. However, the UiB material is not ideally suited in terms of selecting the most common words for children.

4.2.3 Preparation of the stimulus material

The decision was made to record all the monosyllabic words included in the Quist-Hanssen material and the three children's lists selected by Rikshospitalet University Hospital. An experience-based evaluation of these lists concluded that they contained enough words to produce new monosyllabic speech audiometry lists even after removing words judged as unsuitable. The procedure for removing words is presented in section 4.2.5.1. A 62-year-old man with an Eastern Norwegian dialect was chosen to read the material. The same speaker was used for recording the Hagerman material presented in Chapter 2. For the recordings, the speaker was seated alone in an audiometric room, where he read a prepared manuscript of 275 monosyllabic words. The room was equipped with a Norsonic type 1220 microphone, a Norsonic type 1201 preamplifier and a Norsonic type 336 frontend amplifier. The output was routed to line in on a Terratec Phase 24FW soundcard connected with FireWire to a Windows XP personal computer. A/D conversion of the sound was conducted with the maximal 24-bit precision and a sampling frequency of 44 100 Hz. The recording was made with Adobe Audition 1.5 and saved as a standard 32-bit wav file. Matlab routines were developed and used to automatically edit the recorded file and to isolate each of the words and save it as a separate wav file.

4.2.4 Listening tests

4.2.4.1 Second field test: detecting easily and less easily recognizable words

A listening test was prepared to acquire information about the homogeneity of the words. Words that were judged too easy or too difficult to recognize were to be excluded.

The 275 words were normalized to the same equivalent level. The threshold of these words relative to the threshold of the three-word material described in section 3 was estimated. Seven different measurement sets were prepared, each containing a selection of 88-90 out of the 275 words. The words in the measurement sets were used without repetition within a set. Each word was used in three different measurement sets, where one set used them without adjusting their level, the second set used them adjusted +10 dB in level, and finally the third set used them with a -10 dB level adjustment. Combined, the seven measurement sets will test each word with the level adjusted -10 dB, 0 dB and +10 dB relative to the estimated threshold.

Students of the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, administered this test during their practice period in the spring of 2006. Each student was given one of the seven different test sets to perform on up to ten different people. They were instructed to first measure the threshold for the three-word material without noise, and to follow a prescribed adjustment procedure before registering which of the monosyllabic words were recognized. The measurement was performed using standard audiometric equipment in audiometry booths on one ear of a total of 32 subjects with normal hearing. Pooled across all measurements each word was tested 12-18 times overall at the three different levels.

4.2.4.2 First laboratory test: threshold and slope for words measured in hearing level

Students from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, performed this test in the department's laboratory during September 2007. The laboratory test incorporated test items from the Quist-Hanssen speech audiometry, three-word utterances, monosyllabic words, five-word sentences and three-numerals lists.

The test was performed using the finalized lists of the selected monosyllabic words, which were mixed and adjusted in levels as will be described in section 4.2.5.

The main objective of testing the monosyllabic words was to check the hearing threshold and the slope of these words without background noise. The speech audiometers were calibrated before the testing and the results can be presented in hearing level (dB HL).

An example of the scoring protocol can be found in Appendix D. Five test sets (A, B, C, D and E) were made, with the same structure but containing different sentences and words.

The students were instructed to use the scoring protocol in Appendix D, which included three 50-word lists of monosyllabic words (cf. page 5 in the scoring protocol). Points on the performance-intensity curve were to be measured for 10-word groups at 5 dB intervals for one ear. 19 young subjects with normal hearing participated in this test.

A logistic function was estimated for each subject and the median value was used as the slope of monosyllabic words.

In order to determine the calibration level, all the measurements involving 10-word groups were used. These results were pooled together and a logistic function was estimated to determine the speech recognition threshold, forming the basis for the calibration. These measurements could then be compared with similar measurements for three-word utterances on page 2 and digit triplets on page 9 of the scoring protocol (Appendix D). The logistic functions were estimated with the method of least squares by using the solver which is a standard add-in in Excel. The squared errors where weighted with the number of measurements at each level for this calculation.

For this test a calibration signal 1 dB lower than the equivalent levels measured for the speech material without using a weighting filter had been calibrated to 20 dB SPL in the earphones of the audiometer. As described in section 2.2.3.6 the calibration level was corrected for the final material.

4.2.4.3 Second laboratory test: masking level

This test was performed by the author on students and staff members from the Audiology Programme at the Faculty of Health Education and Social Work, Sør-Trøndelag University College, in the laboratory of the department during April 2008. The laboratory test incorporated five-word sentences, three-word utterances and monosyllabic words. The measurements were performed using a clinical audiometer calibrated for HiST speech audiometry, and one ear was selected for each test subject.

The speech noise of the audiometer was mixed ipsilaterally with the speech signals, and presented to the ear chosen by the test subject. Five subjects with normal hearing participated in the test. The scores for 10-word groups without noise were obtained first. Speech noise was then introduced, and measurements were performed as signal to noise ratios were decreased in 5 dB steps until a score of 0 % was obtained. The signal-to-noise ratio for a 50 % score was estimated by fitting a logistic curve through a maximum likelihood procedure for each subject. The mean value of the estimated signal-to-noise ratios among the subjects was calculated. The level of the speech noise was held constant after it was introduced, and the level of the speech material was varied. The level of the speech noise from the audiometer (GN Otometrics Aurical Plus) was measured to 71.3 dB SPL without frequency weighting from the supraaural earphones TDH 39 with MX-41/AR cushions mounted on an IEC 60318-3 (1998) acoustic coupler. The spectrum of the speech noise was not measured.

4.2.5 Preparation of the final material

4.2.5.1 Word inclusions and exclusions

Out of the 275 recorded words a selection had to be made for the new monosyllabic speech audiometry lists.

Some of the words had been recorded more than once, and the best recording of replications was selected by listening.

Table E.1 in Appendix E shows the list of all the monosyllabic words. Column 2 shows all the words that were evaluated. The words selected for the new monosyllabic speech audiometry lists are printed in bold font. To assist in the process of selecting words several columns were included in this table.

Columns labelled *Included in Oxford 3000* and *Not Included in Oxford 3000* show a translation of the Norwegian words into English. The Oxford 3000 (Hornby 2005) is a selection of 3000 keywords in English. Language experts and teachers made this careful selection of words which should receive priority in vocabulary study, choosing them on the basis of their importance and usefulness. The selection is based on the following three criteria. First, *"the words which occur most* ***frequently*** *in English are included based on The British National Corpus and the Oxford Corpus Collection"* (Hornby 2005, p.R99). Second, *"only those words which are frequent across a* ***range*** *of different types of text"* are included (ibid.). Finally, *"the list includes some very important words which happen not to be used frequently, even though they are very* ***familiar*** *to most users of English"* (ibid.). The ideal situation for our purposes would be to have such a list not for English words but for Norwegian words. However, similar information for Norwegian words was not found, and the decision was therefore made that this list in English could be used as one factor in the process of selecting words. If after translation of the Norwegian words into English the word was not found on the Oxford 3000 wordlist, the translated word was put in the Not included in *Oxford 3000 column* in Table E.1. This was used as an indicator that the word might be excluded from our own final list.

The column labelled *Norwegian Google pages 2006-09-20* gives an indication of how frequent the words are. In section 4.2.2 the frequencies of the different words are evaluated based on the UiB material. To get another indication of word frequencies number of Norwegian Google pages containing each word was recorded on a single date. This information is presented in the column *Norwegian Google pages 2006-09-20.* A low score in this column is used as an indicator that the word may not be suitable for the new selection of monosyllabic words. Our reason for using Google and not the UiB material for this evaluation was that Google represented a larger text base, and was supposed to be more up-to-date for words used in Norwegian today.

The columns *Number of tests in listening test* and *Per cent recognized* show the results of the listening test described in section 4.2.4.1. The first of these two columns presents the number of tests conducted with each word in total, with the level adjusted by -10 dB, 0 dB or +10 dB relative to the estimated threshold for monosyllabic words. The second of these two columns gives the results from the listening tests in per cent recognized. If the score of a word was very high or very low, this was used as an indicator that the word might be excluded. Words not excluded from the new list but scoring lower than 39 % or higher than 77 % are marked D or E, respectively, in the column *D = difficult to recognize E = easy to recognize.* This information was used as described in 4.2.5.3 when mixing the words: we ensured that the easy and difficult words were spread over the different lists.

The four last columns show the lists the words originated from and finally whether or not the word is to be excluded in the new selection of monosyllabic words. The selected words are also shown in bold font in the word column.

Of the 275 recorded words a selection of 160 words was made on the basis of Table E.1. It was decided to include all the words in Rikshospitalet's words for children in our selected 160 words, even if other indicators in the table indicated exclusion. The basis for this decision was that the words have been judged as appropriate for children by experts in speech audiometry for small children. On the other hand, several of the Quist-Hanssen words were excluded on the basis that they were lacking in Oxford 3000, present on only few Norwegian Google pages, that they achieved a high or low score in the listening test, or due to a combination of these factors. The excluded words are marked with an "X" in the *Excluded* column and the factors feeding into the decision to exclude them are marked in bold font in the appropriate columns.

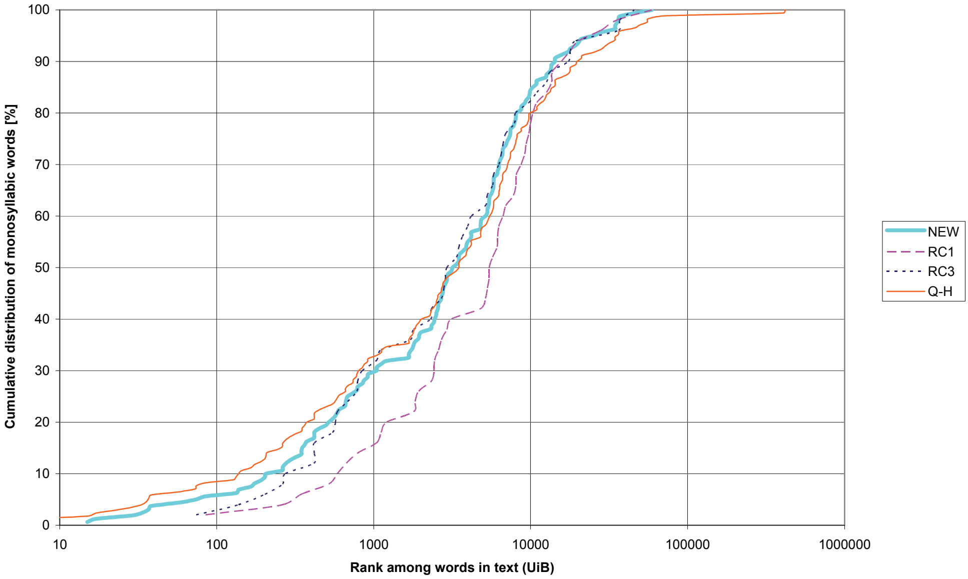


Figure 4.3 The cumulative distribution of our new monosyllabic words (NEW), Rikshospitalet's monosyllabic lists for children (RC1 and RC3) and Quist-Hanssen's (Q-H) monosyllabic words as a function of their ranking among the UiB words.

Figure 4.3 shows the cumulative distribution of the final selection of monosyllabic words (NEW) compared to RC1, RC3 and Quist-Hanssen's words (Q-H) ranked according to the UiB material. Our new selection of words has a slightly delayed rise compared to Quist-Hanssen's words. The reason for this delay can be found in the decision to include RC1 and RC3 in this new material. Both of these lists incorporate words which are less frequent according to the UiB material. As stated in section 4.2.2.2 the UiB material is not ideal in terms of identifying the most common words for children. The figure shows that the new word material has a steeper rise when approaching 100%. This new selection should therefore represent a selection of words which are more common today than the selection made by Quist-Hanssen.

4.2.5.2 Phonemic balance

Figure 4.4 shows the phonemic distribution of the new selection of monosyllabic words compared to the estimated distribution in Norwegian text, estimated as described in section 2.2.1.3. Using only 160 monosyllabic words it is hard to achieve a better balance than what is shown in Figure 4.4. The greatest difference is found in the distribution of the schwa sound, as has to be expected for monosyllabic words.

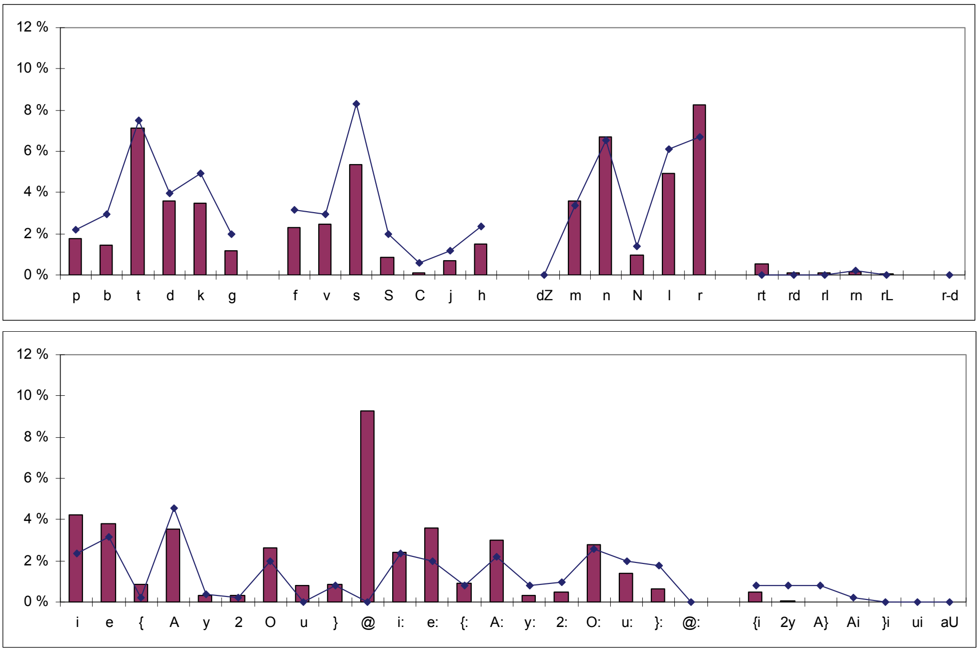


Figure 4.4 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The line with diamonds shows the distribution of the new selection of monosyllabic words.

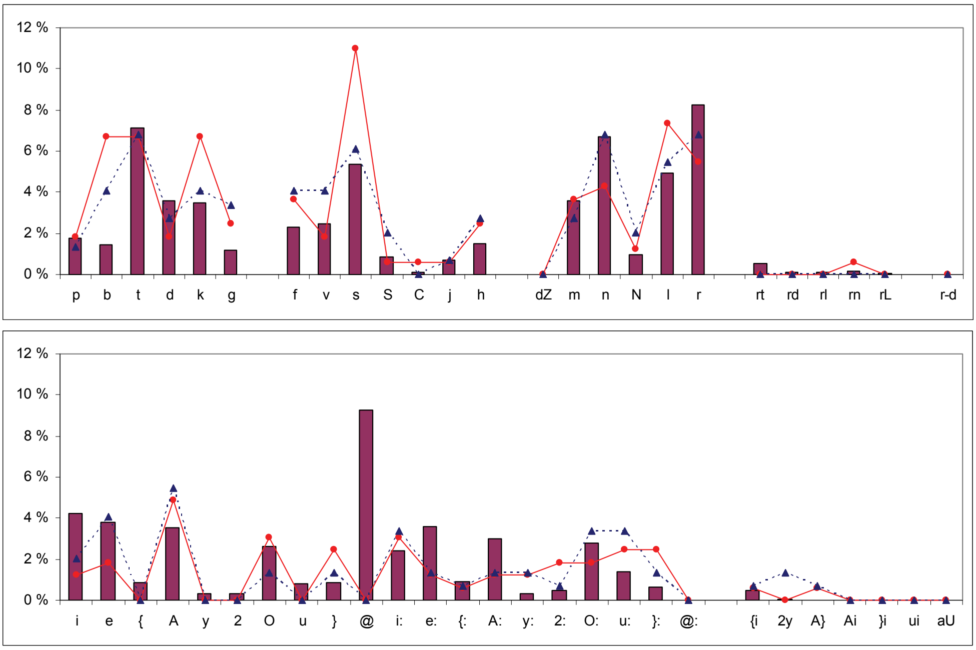


Figure 4.5 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The solid line with circles shows the distribution for RC1 monosyllabic words and the dotted line with triangles shows the distribution for RC3 monosyllabic words.

Figure 4.5 shows a comparison between the phonemic distribution of the Rikshospitalet's words for children, lists RC1 and RC3, and the estimated distribution in Norwegian text. As expected, the differences here are greater because each list contains only 50 words.

In Figure 4.4 we find that for the new material of 160 monosyllabic words only one phoneme (s) differs by more than 2 percentage points from the columns if we exclude schwa. In Figure 4.5 we find that for the RC1 list 8 phonemes (b, k, s, n, l, r, i and e:) differ by more than 2 percentage points from the columns. For the RC3 list 5 phonemes (b, g, i, e: and u:) differ by more than 2 percentage points from the columns.

4.2.5.3 Mixing strategies

Quist-Hanssen's monosyllabic words have had a prominent place in Norwegian speech audiometry. Some institutions have neglected to use spondee measurements and have performed speech audiometry only with the monosyllabic words. One of the reasons for this is that the measurements have had good reproducibility. Traditionally, the measurements have been obtained by using only 10 words at each level. The measurements have usually started at a random position in the list of words, and have been performed using the following words organized as groups of ten. This procedure has functioned rather well because Quist-Hanssen developed the material in such a way that it displayed a good mixture of the words.

A mixing procedure was developed in Matlab in order to try to achieve some of the good qualities found in Quist-Hanssen's lists with the new monosyllabic words. We had our new selection of 160 monosyllabic words and the goal of the mixing procedure was to generate a list containing 480 words, made up from the 160 words by using these three times but with a new order for each repetition. The mixing procedure was based on the following principles:

1. A group of 50 words was drawn up at random from the 160 different words available. These words were only used for initiating the procedure and were discarded later.

2. For each of the words available for selection a score was generated for the combination of this new word and the 9 as well as the 49 words last selected. The word with the lowest score was then selected as the next word. The score was calculated from 6 different parameters (A, B, C, D, E and F in Equation 4.1) as presented in point 3 below.

3. Parameter A was based on the phonemic distribution in the potential 10-word group of 52 individual phonemes shown in Figure 4.4.

Parameter B was based on the phonemic distribution in the potential 10-word group grouped together in 6 of the major phoneme groups also shown in Figure 4.4 (group1: p, b, t, d, k and g; group2: f, v, s, S, C, j, h and dZ; group3: m, n, N, l and r; group4: I, e, {, A, y, 2, O, u and }; group5: i:, e:, {:, A:, y:, 2:, O:, u: and }:; group6: {I, 2I, A}, Ai, }I, ui and aU).

Parameter C was based on the phonemic distribution score of the potential 50-word group for the 52 individual phonemes.

Parameter D was based on information about the number of difficult words in the potential 10-word group.

Parameter E was based on information about the number of easy words in the potential 10-word group.

Finally, parameter F was based on the requirement that a word should not be reused before 50 other words had been selected.

All the parameters (A, B, C, D, E and F) were based on the principle that a perfect fit would give the number 0 for a parameter. A perfect fit for a given parameter meant that the distribution in the sub- sample (10 or 50 potential words) was the same as the distribution in the list of 160 monosyllabic words for the quality (phonemes, phoneme groups etc.) under evaluation. A greater difference from a perfect fit would give a larger parameter. An example of the calculations can be given for parameter C. Parameter C was based on the phonemic distribution score of the potential 50-word group for the 52 individual phonemes, but in this example we show the calculation for the two first phonemes only. Phoneme 1 has a frequency of 11 and phoneme 2 has a frequency of 15 in the 160 different words. The optimal number of these phonemes in a 50 word list should be $11 \cdot 50/160 = 3.4375$ and $15 \cdot 50/160 = 4.6875$ respectively. If a combination of a potential new word and the 49 most recent selected words gives a frequency of 5 for phoneme 1 and a frequency of 4 for phoneme 2, parameter C should be calculated as: $C = \sqrt{(5-3.4375)^2 + (4-4.6875)^2}$. In the deployed routine the equation was of course expanded with parts for all of the 52 phonemes that were evaluated, and not only the two parts shown in this example .

The final score for each potential word was computed as the multidimensional Euclidean distance.

$$score = \sqrt{A^2 + B^2 + C^2 + 100 \cdot D^2 + 100 \cdot E^2 + 100 \cdot F^2} \tag{4.1}$$

The factor 100 was incorporated to ensure that there would always be a perfect balance of one easy and one difficult word in each 10-word group, and that a word would never be repeated until 50 different words had been selected.

4. Points 2 and 3 were repeated 25 times for the first 160 word selections, and the list with the best score for the final word selected was saved.

5. The procedure continued with the selection of words 161-480 as described in point 3.

6. If the produced list had multiple occurrences of identical word pairs the procedure started again from point 1.

7. When the procedure was completed, the resulting list was exported to Excel for further processing.

The Matlab mixing procedure was run and after 44 repetitions the procedure identified an accepted list. The first 450 word in this list are presented as nine 50-words lists in Appendix E. The scores for all of the 441 possible combinations of 10 consecutive words are presented in Figure 4.6. The two peaks in the figure close to groups 160 and 320 are due to the few alternative words available for selection when the procedure has selected almost all of the 160 words that must be used before restarting the selection process.

Figure 4.7 presents the phonemic distribution for each of the nine 50-word lists in Appendix E compared with the estimated distribution in Norwegian text. The balance between the lists is reasonably good, and it is difficult to achieve better results because of the limited amount of phonemes available in lists of this size.

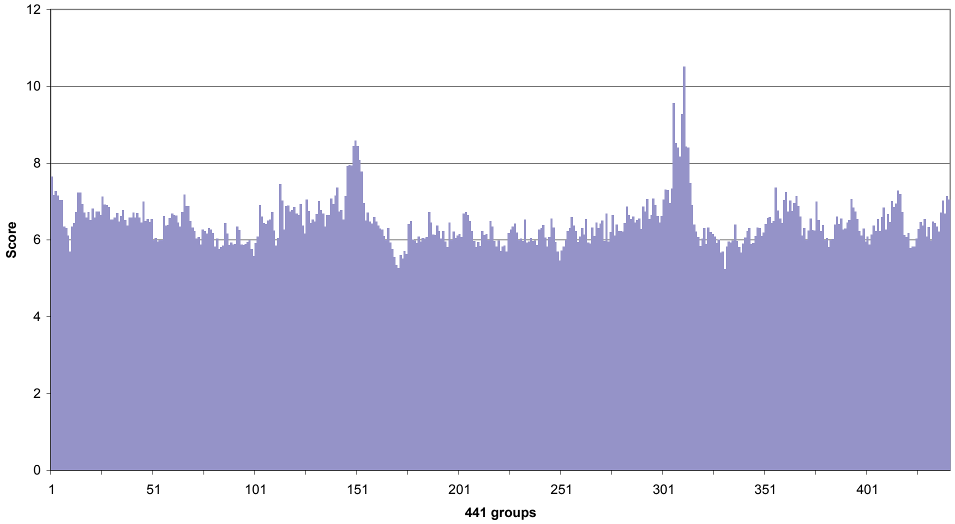


Figure 4.6 Score for 10-word groups of monosyllabic words in Appendix C.

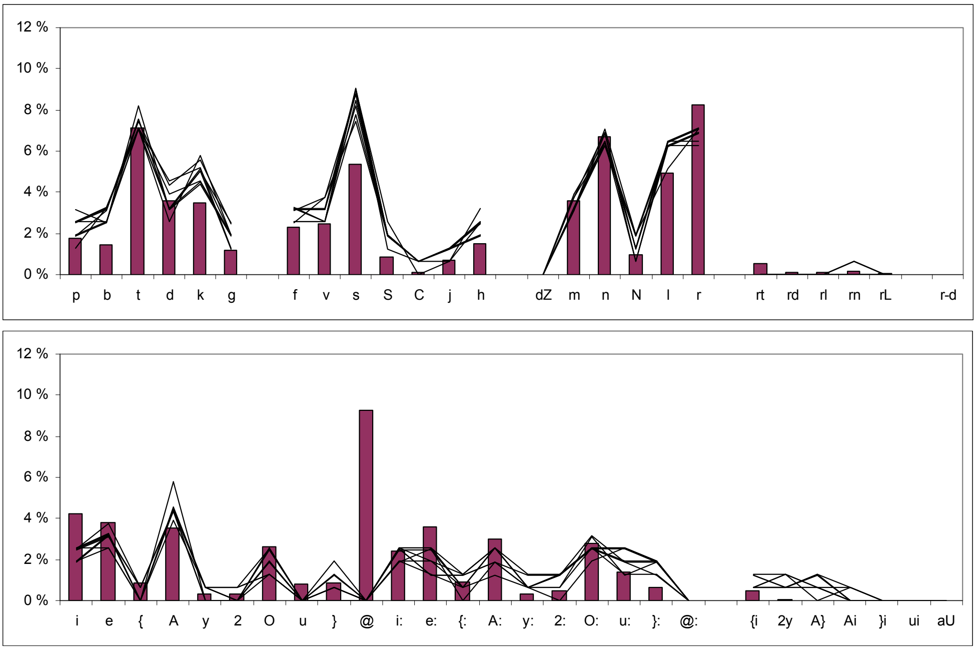


Figure 4.7 The columns show the distribution of phonemes for the 20 000 most frequent words in the University of Bergen (UiB) material, corrected for the frequency of each word. The solid lines show the distribution of 9 different lists of 50 monosyllabic words.

4.2.5.4 Level adjustments

The monosyllabic words will primarily be used for measurement of maximum speech recognition score. This type of measurement could require different level adjustments from the materials presented in Chapters 2 and 3 which were designed for threshold measurements. The materials used for threshold measurements were chosen on the basis of the requirement that all the words must have approximately the same threshold either with or without background noise. For measuring maximum speech discrimination the requirement can be that the words have approximately the same loudness at loud levels. Three different level adjustment strategies were evaluated by informal listening tests, namely unadjusted, equivalent level and loudness normalized.

4.2.5.4.1 Unadjusted

The unadjusted material comprises the sound files with the natural levels from the recording. The monosyllabic words were recorded as described in section 4.2.3 and processed with a Matlab routine which isolated each word and saved it as a separate wave file.

4.2.5.4.2 Equivalent level normalized

Before the listening test described in section 4.2.4.1 the words were normalized to the same equivalent level. A Matlab procedure fetched all the wave files described in section 4.2.5.4.1 containing an unadjusted word. The procedure calculated the linear equivalent level of each of the unadjusted words. This data was transferred to an Excel spreadsheet for further processing to determine the exact adjustment necessary for each word. The mean equivalent level of all the words was -21.6 dB relative to the clipping limit of the wave file. To avoid clipping, all words were adjusted to -22.6 dB equivalent level relative to clipping. To achieve this, the words had to be adjusted by between -5.9 and 5.4 dB. Another Matlab procedure was employed to fetch all the wave files described in section 4.2.5.4.1 containing an unadjusted word, perform the required level adjustment, and store the level-adjusted word in a wave file with a new name.

4.2.5.4.3 Loudness normalized

The monosyllabic words are intended mainly for measuring maximum discrimination. This is a way of measuring supra threshold and requires rather strong stimuli. In order to eliminate level differences between the words we decided to normalize all the words to the same loudness.

Hastings (2002) has developed Matlab code for calculating loudness according to ISO 532 (1975) Method B. The Matlab code is based on the BASIC code of Zwicker et al. (1984). Timoney et al. (2004) present tests of this and four other Matlab implementations for calculating loudness. One conclusion was that none of the implementations produced the exact figure for loudness but that all are rather close to the expected value.

A limitation of the ISO 532 Method B standard is that it is only specified for calculating the loudness of steady sounds. Glasberg and Moore (2002) have proposed a model of loudness applicable to time-varying sounds, but standards for this have not been established yet and we were unable to find code for using this model. Zwicker (1977) describes a model for loudness of temporally variable sounds and evaluates it for several types of sounds that vary both temporally and spectrally. For connected speech the conclusion is that the hearing system perceives speech such that the loudest parts of a spoken sentences are responsible for loudness. The ISO 532 Method B standard calculates the loudness on the basis of the spectrum of the sound. Hastings' routine is based on a function in the Signal Processing Toolbox of Matlab called PSD-Power Spectral Density Estimate. In Hastings' routine this function is utilized without overlapping for estimating the spectrum. Steady-state sounds will be correctly estimated with Hastings' method, but since we are interested in the correct spectrum for such transients as monosyllabic words, the code was altered and overlapping of 0.99 of the window size was introduced. In the altered model the spectrum will be correctly calculated for short sounds by the Matlab code, and the length of the monosyllabic words does not vary to any great extent. Against the background of these facts normalization based on loudness calculation was chosen as the level adjustment procedure for the monosyllabic words even if the ideal method could not be used. While performing the loudness calculations we decided to use the diffuse sound field correction, because earphones will be used for most of the measurements involving these words.

4.2.5.4.4 Level adjustments, results

The loudness of the monosyllabic words adjusted according to the three methods described in section 4.2.5.4.1-4.2.5.4.3 was evaluated with loudness measurements based on ISO 532 Method B as described in 4.2.5.4.3. The results for the 275 words recorded are presented in Figures 4.8-4.10.

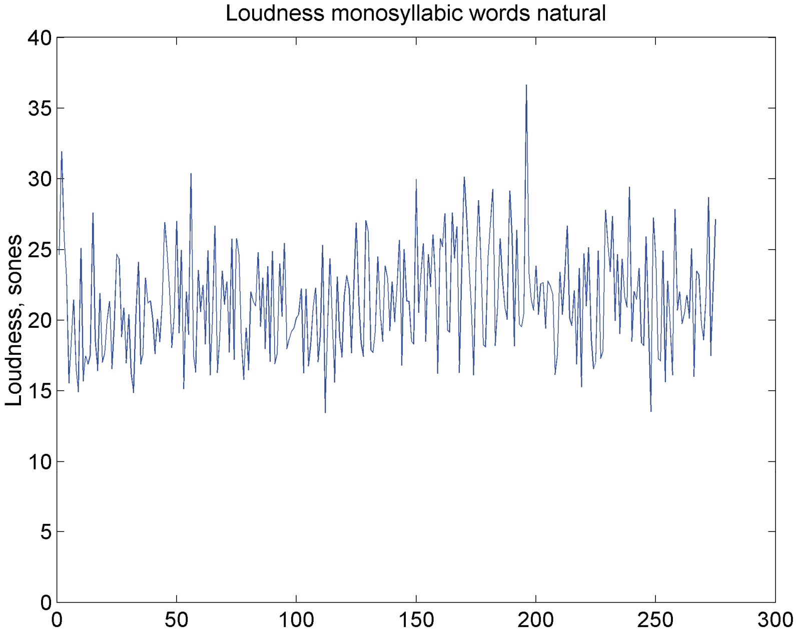


Figure 4.8 Loudness levels estimated for the unadjusted monosyllabic words described in section 4.2.5.4.1.

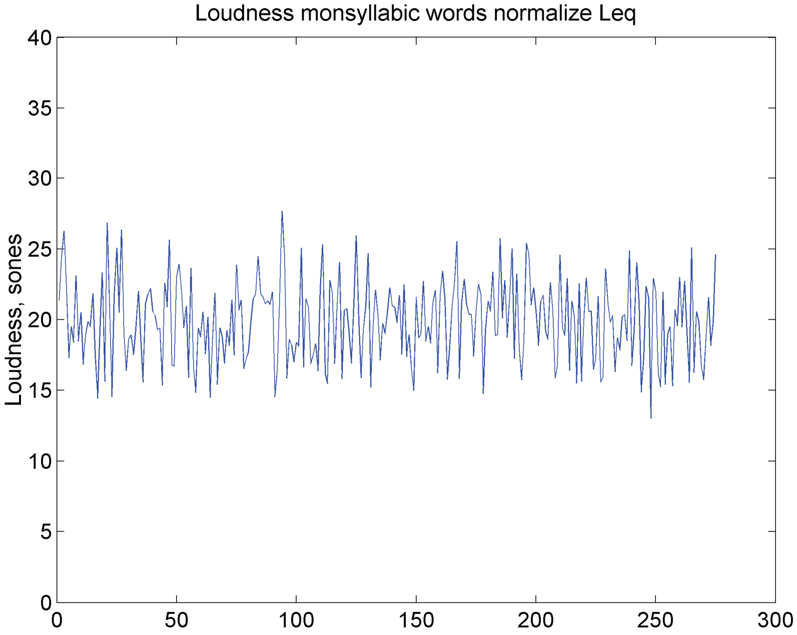


Figure 4.9 Loudness levels estimated for the monosyllabic words normalized to the same equivalent level as described in section 4.2.5.4.2.

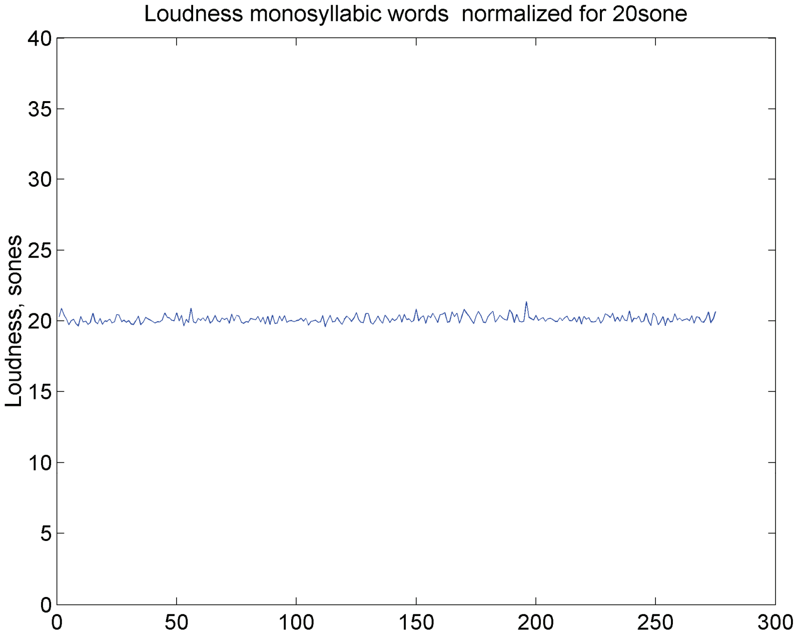


Figure 4.10 Loudness levels estimated for the monosyllabic words normalized to the same loudness (20 sones) as described in section 4.2.5.4.3.

Figure 4.10 shows the almost ideal result as expected because we are verifying the results with the same tools as the ones used in producing them, but informal listening tests also confirm that the loudness normalization gives a balanced impression when performed at loud levels. The lists where the equivalent level had been normalized were the least natural-sounding ones. We therefore decided to make the final lists with the levels adjusted according to the loudness normalized procedure.

4.3 Results

4.3.1 Second field test: detecting easily and less easily recognizable words

The results from this test are given in Appendix E Monosyllabic words, Table E.1. For each word we find how may times a word was tested in total at the -10, 0 and +10 dB levels. The result is shown in the column Number of tests in the listening test. The percentage recognized words is presented in the column Per cent recognized. The results were used as an indicator for inclusion or exclusion in the material (section 4.2.5.1), and to determine the most easy and most difficult of the included words in terms of their recognition so that they could be spread evenly across the lists of monosyllabic words (section 4.2.5.3).

4.3.2 First laboratory test: threshold and slope for words measured in hearing level

The results were analyzed using the methods described in section 4.2.4.2. The median value of the slope was 7 %/dB. The value of the speech recognition threshold was 8.8 dB HL. As reported in section 2.2.3.6 the calibration level has later been changed.

4.3.3 Second laboratory test: masking level

We found that to calibrate the speech noise of the audiometer used (GN Otometrics Aurical Plus) in terms of effective masking level, the level of the speech noise should be 25 dB over the attenuator setting when measured without frequency weighting on the supraaural earphones TDH 39 with MX-41/AR cushions mounted on an IEC 60318-3 (1998) acoustic coupler.

4.4 Discussion

The results from the second field test were helpful in terms of weeding out the words that were the easiest and the most difficult ones to recognize. 22 words were rejected because they were difficult to recognize, but only one because it was easily recognizable. Of the 160 remaining words, 16 with a score under 39 % were marked with a D for difficult to recognize, and 16 with a score over 77 % were marked with an E for easily recognizable. These marks were used in the mixing procedure, so that only one D-word and one E-word are represented in each 10-word group in the lists.

The results from the second laboratory test were helpful in establishing the speech recognition threshold and the slope for the monosyllabic words. The results can be seen in Figure 6.1. It is apparent that monosyllabic words have the highest threshold among the different word types and the shallowest slope. This behaviour is as expected for monosyllabic words.

4.5 Conclusion

In this chapter a new speech audiometry test using monosyllabic words for measurement of maximum speech recognition score has been described. The selected words are in common use and the lists have a good phonemic balance. The level is adjusted to the same theoretical loudness for each word. 160 words were selected and these are repeated in different order to make up a total of nine lists containing 50 words each. The words have been mixed to produce theoretically equivalent lists.

Performance intensity curves have been established for subjects with normal hearing. The equivalence of these lists and performance intensity curves for hearing impaired persons remains to be studied.

Recordings of lists of monosyllabic words selected for children have been made available.

Chapter 5

Evaluation of measurement strategies

5.1 Introduction

In speech audiometry, the speech recognition threshold is usually defined as the level where we measure a 50 % score. The goal when performing speech audiometry is often to measure this point with reasonable accuracy. For suprathreshold measurements of monosyllabic words it can also be of interest to measure maximum speech recognition score and rollover index. The slope of the performance-intensity curve is another parameter that one may need to estimate. One way of estimating these parameters can be to measure points on the performance-intensity curve and then use the curve to estimate the desired parameters. Several procedures also exist for simply measuring the desired parameter directly.

In this chapter several procedures for estimating the threshold and some of the other parameters are evaluated by simulations. In section 5.1.1 a representation of the performance-intensity function is presented, and four hypothetical subjects are defined. These will be used in the simulations to evaluate how good the different measuring procedures are in terms of estimating the correct answer. Section 5.1.2 presents the statistics of the binomial distribution, which describes the statistics of speech audiometry. Section 5.2 describes context effects that influence the measurement involving five-word sentences and three-word utterances. Section 5.3 presents the methods used for the simulations. Section 5.4 gives the results of the simulations. In the discussions of section 5.5 the results regarding the performance of the different procedures are evaluated and recommended procedures for “HiST taleaudiometri” are suggested.

The levels in dB without a reference level presented for the simulations in this chapter are given like that to be universal. They may represent measurements performed in dB HL, dB SL, dB SPL or dB SNR, although the motivation for the selected range is dB SPL.

5.1.1 The performance-intensity function

In the preceding chapters we have used the sigmoid function as a special case of the logistic function (cf. Equation 2.1) to fit performance-intensity curves to our measurement data. This function describes the results well when the score varies from 0 % to 100 % with increasing levels or signal-to-noise ratios, as achieved in our testing of the speech audiometry material using young subjects with normal hearing. When performing speech audiometry on persons with hearing disorders the maximum recognition score scores will not always reach 100 %. Kollmeier et al. (2008) give a more general logistic function for fitting speech intelligibility (*SI*) as a function of speech level *L* to empirical data:

$$SI\,(L) = SI_{\max}\,\frac{1}{1 + \exp\left(-\dfrac{L - L_{mid}}{s}\right)} \tag{5.1}$$

and the slope at the midpoint is given by the following equation

$$slope = \frac{SI_{\max}}{4s} \tag{5.2}$$

where $SI_{\max}$ is a parameter for maximum intelligibility, L_{mid} is speech level of the midpoint of the intelligibility function and *s* is a slope parameter.

For some hearing disorders the performance-intensity curve does not continue to rise with increasing levels when once loud levels have been reached. We call this effect rollover and have modified the equations above to include this effect:

$$SI(L) = RO(L)\cdot SI_{\max}\,\frac{1}{1+\exp\left(-\dfrac{L-L_{mid}}{s}\right)} \tag{5.3}$$

where we have introduced a level dependent variable *RO*(*L*) for rollover given by:

$$\begin{aligned} &RO(L) = 1 && \text{when } L \le 60 \\ &RO(L) = 1 - rof\,\frac{(L-60)^2}{20^2} && \text{when } 60 < L \le 80 \end{aligned} \tag{5.4}$$

rof is a parameter between 0 and 1 giving the degree of rollover (rof= 0 means no rollover, *rof*= 1 means full rollover at a level of 80 dB).

Equations (5.3) and (5.4) were designed for these simulations and are not to be used as a general function fitting the performance-intensity curves for all types of hearing loss. The rollover function in particular is only introduced to simulate one variant of rollover limited to levels in the range between 60 and 80 dB.

Equations (5.2)-(5.4) will be used in this chapter to simulate performance-intensity curves. We will simulate the behaviour of four hypothetical subjects (HS1-HS4), whose speech audiometry functions are described by means of different performance-intensity curves. The parameters for these hypothetical subjects are given in Table 5.1.

Table 5.1 Parameters used to simulate the four hypothetical subjects.

Parameters in equations (5.2)-(5.4) Hypothetical subject	L_{mid} (threshold) [dB]	*slope* [%/dB]	$SI_{\max}$ (maximum speech intelli-gibility) [%]	*rof* (rollover factor)
HS1	35	10	100	0
HS2	35	3	100	0
HS3	35	4	80	0
HS4	35	4	80	0.5

The performance-intensity curves for these four hypothetical subjects are shown in Figure 5.1. The vertical line at the 35 dB level shows the thresholds for the midpoints at 50 % score for HS1 and HS2, and 40 % score for HS3 and HS4 marked by the horizontal lines. The decision was made to keep the threshold constant at 35 dB for our hypothetical subjects in order to evaluate how the different shapes of the performance-intensity functions influenced the results. Subjects with different hearing thresholds can have similar performance-intensity curves shifted horizontally. Speech audiometry could be performed on these subjects by adjusting the range of test levels used in an appropriate way to achieve the desired results. The starting level for the simulations is usually selected close to the expected threshold. To avoid influences on the results from tying up the starting level and the threshold, the starting level was usually randomized within ±5 dB. The parameter values were chosen to simulate a representative selection of different types of hearing. HS1 could represent a subject with normal hearing, a light sensorineural hearing loss or a conductive hearing loss. The slope of 10 %/dB at the midpoint of the performance-intensity function is close to what we expect when performing speech audiometry procedures with three-word utterances or five-word sentences without noise. HS2-HS4 represent variants of sensorineural hearing loss. HS2 has a very low slope on the performance-intensity function, HS3 has a low slope and reduced maximum speech recognition score and HS4 has the same characteristics as HS3 but includes rollover, which reduces the intelligibility for loud speech. This could be a subject with a retrocochlear hearing loss.

In section 5.4 we will give the results from simulations of speech audiometric procedures performed on these four hypothetical subjects. Different speech audiometric procedures are evaluated. All of the procedures produce an estimate of the threshold and some of the procedures also give estimates for some of the other parameters. All the estimated parameters do not need to be defined exactly like equations (5.2)-(5.4) and the correct estimate for the particular procedures will be described in sections 5.3.1-5.3.6. By repeating each procedure several times, statistics of the measurements can be obtained.

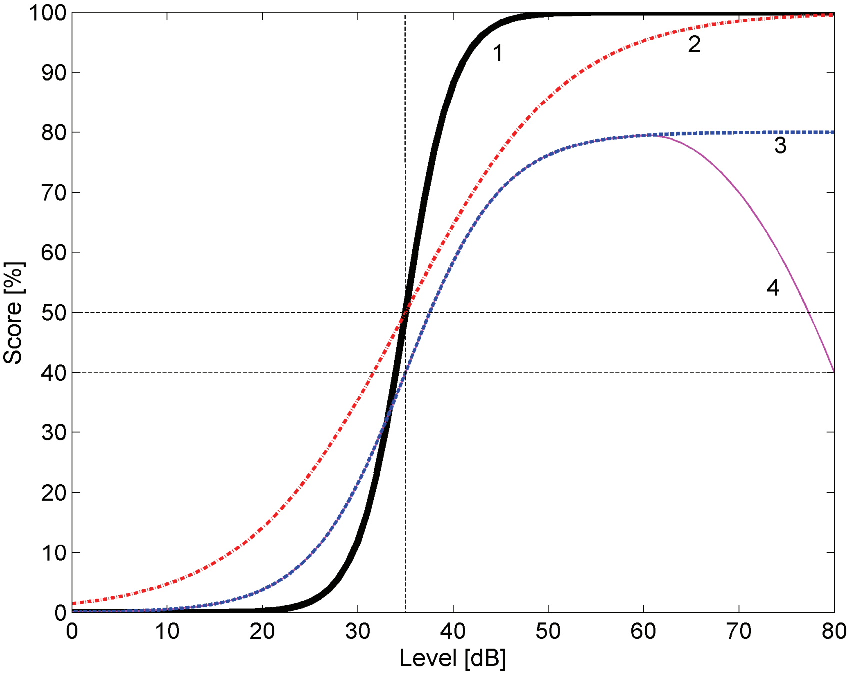


Figure 5.1 Performance-intensity curves for four hypothetical subjects (HS1-HS4).

The slope of the performance-intensity curves is not only affected by the degree of the hearing loss, but different types of speech material will give different slopes for the same user. For the speech audiometry material in “HiST taleaudiometri” we have found in Chapters 2-5 and 6 the following slopes for young listeners with normal hearing when measured around the speech recognition threshold without noise: The monosyllabic numerals, 17 %/dB; three-word utterances and five-word sentences, 10 %/dB; and monosyllabic words, 7 %/dB. When measuring with noise the slopes are found to be 16 %/dB for the three-word utterances and 14 %/dB for the five-word sentences. In a real test situation the resulting slope of the performance-intensity function will be a combination of the specific speech material and the listener's performance with this material.

We decided to perform the simulations with our four hypothetical subjects without considering the effects of using different speech materials. This allows us to assess the accuracies of different measurement procedures, and we can use some of our hypothetical subjects to simulate a kind of worst case scenarios. After having discussed the actual measurement procedures and finally decided upon one specific method to recommend, we need to consider how accurate the expected results will be with this measurement procedure and the selected speech material on a typical subject. In order to estimate this accuracy we should have knowledge of the slope of the speech material when used on the individual subject, but we lack this information. Moreover, we also lack evaluations of how "HiST taleaudiometri" functions with different types of hearing loss. In the literature some studies can be found with comparisons of the slope of the performance-intensity curve for subjects with normal hearing and subjects with hearing loss. Cooper and Cutts (1971) compared normal-hearing and sensorineurally impaired subjects with Northwestern University Auditory Test No. 6 (open set monosyllables) in cafeteria noise. The slopes for the two groups were not significantly different, at 3.57 %/dB for normal-hearing versus 3.47 %/dB for hearing-impaired subjects. Gang (1976) measured speech recognition threshold without noise with CID Auditory Test W-22 (open set monosyllables) on subjects with presbycusic hearing loss. The slope was 1.4 %/dB compared to the norm of 6 %/dB. Beattie and Warren (1983) investigated slope with CID W-22 without noise on subjects with mild to moderate sensorineural hearing loss. The slopes remained at approximately 3 %/dB when the thresholds varied over a range of 45 dB. For audiograms that progressed from flat to steeply falling, the slopes decreased from about 3.5 %/dB to 2.5 %/dB. Beattie and Raffin (1985) used CID W-22 without noise and checked slope among both normal-hearing listeners and subjects with mild-to-moderate sensorineural hearing loss. The slopes were 4.9 %/dB for normal-hearing subjects and fell to 2.7 % for the hearing-impaired. Beattie (1989) measured word recognition functions with CID W-22 in multitalker noise on subjects with normal hearing and with mild-to-moderate hearing loss. The slope for the hearing-impaired group was 2.6 %/dB compared to 3.6 %/dB for the normal-hearing subjects. Wagener (2003) tested the slope of the performance-intensity function both of hearing-impaired and normal-hearing subjects with the five-word Oldenburg sentences with several types of noise, both stationary and modulated. These sentences are the German equivalent of our five-word sentences. Wagener found that the slope values hardly differed between the hearing-impaired and the normal-hearing subjects, with median slopes of 14.9 %/dB and 17.3 %/dB respectively.

For the most part in these results we find first that when testing with stationary or modulated noise the slopes of the performance-intensity curves are almost identical for hearing-impaired and normal-hearing subjects. In contrast, we also find that when testing with monosyllabics without noise the slopes are highest for normal-hearing subjects and decrease with hearing loss, and that the decrement seems larger for steeply falling audiograms. Nevertheless, we still lack information about how the test materials with higher slopes, such as five-word sentences, three-word utterances and monosyllabic numerals behave without noise for subjects with or without hearing losses.

Generally speaking, we can use the simulations with hypothetical subjects HS1-HS4 as a means to assess how different measurement procedures perform for different subjects with monosyllabic words without noise. For the speech recognition thresholds HS1 will give close to expected results, but with some uncertainty because of the effect of a possible hearing loss. For testing with monosyllabic numerals without noise or three-word utterances and five-word sentences with noise we expect a higher slope, and some extra simulations with a suitable hypothetical subject may be needed.

5.1.2 Binomial statistics

If we could perform speech audiometry at a specified level with a list containing an infinite number of ideal test words we would obtain the correct score for this measurement, 100*p* %. The parameter *p* denotes the probability of obtaining a correct response for words tested at this level. When performing real speech audiometry with repeated lists, each containing a finite number of test words, the results will scatter around the correct score. The statistics describing the performance of speech audiometry tests if all the words have the same probability is the binomial distribution (Carney and Schlauch 2007; Hagerman 1976; Gelfand 2003; Gutnick and St. John 1982; Lyregaard 1997; Raffin and Schafer 1980; Thornton and Raffin 1978).

The probability of getting exactly *k* successes when performing speech audiometry with test lists containing *n* words, at a level where *p* is the probability of getting a correct response, is given by the probability mass distribution *f* which for *k* following the binomial distribution can be written as follows:

$$f(k;n,p)=\binom{n}{k}p^k(1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots, n$ and where (5.5)

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

is the binomial coefficient. The standard deviation can be calculated by:

$$\sigma = 100\sqrt{\frac{p(1-p)}{n}} \quad [\%] \tag{5.6}$$

Thornton and Raffin (1978) found excellent agreement between the prediction by Equation 5.6 and the standard deviations of repeated speech scores for thousands of hearing-impaired subjects. The standard deviations are given in Table 5.2 for some values of p and n. It can be observed that when the expected score is close to 50 % a very large number of items is needed in the list in order to obtain a low standard deviation, whereas when the expected score is close to 100 % or 0 % just a few items are needed in order to obtain comparable standard deviation. To give an example, a 1000-item list gives a standard deviation of 1.6 % when the expected score is 50 %, whereas we only need a 5-item list in order to get a better standard deviation of 1.4 % when the expected score is 99.9 % or 0.1 %.

Table 5.2 The standard deviation of a list score with various items per list for different expected true scores. A binomial distribution is assumed.

Expected score (100*p*) Number of items in test list (*n*)	50 %	90 % or 10 %	99 % or 1 %	99.9 % or 0.1 %
3	29 %	17 %	5.7 %	1.8 %
5	22 %	13 %	4.4 %	1.4 %
10	16 %	9.5 %	3.1 %	1.0 %
30	9.1 %	5.5 %	1.8 %	0.6 %
50	7.1 %	4.2 %	1.4 %	0.4 %
100	5.0 %	3.0 %	1.0 %	0.3 %
1000	1.6 %	0.9 %	0.3 %	0.1 %

In practice individual words in a list do not have the same probability of being recognized. Lyregaard (1997) states that the simple binomial distribution must be replaced with the more general subnormal binomial distribution to describe the statistics exactly. Then a simple analytical expression is not possible, but as long as the distribution of probabilities is not too wide he concludes that the dispersion will only be slightly less than what we get from the simple binomial.

In our simulations we have assumed that each item in the test list has the same probability of being recognized as a prerequisite in order to use the simple binomial distribution for describing the statistics.

5.2 Context effects on sentence recognition

Boothroyd and Nittrouer (1988) have proposed to use a factor *j* to describe how the probability of recognition of wholes is related to the probability of recognition of the constituent parts. Given that the recognition of a whole sentence requires the recognition of all the words and that each word has the same probability of being recognized, then

$$p_s = (p_w)^n \tag{5.7}$$

gives the relationship between the probability of understanding the sentence, p_s and the probability of understanding the words, p_w where *n* is the number of words in the sentence. If there were no context effects at all we would expect $n = 3$ for three-word utterances and $n = 5$ for five-word sentences. Because the sentences/utterances described in Chapters 2 and 3 are syntactically correct and constructed from a limited number of words we assume that the exponent in Equation 5.7 is lowered and we can thus write the equation

$$p_s = (p_w)^j \tag{5.8}$$

where $1 \leq j \leq n$. If we needed to recognize all the words to recognize the sentence then $j = n$. If $j = 1$ the recognition of any word in the sentences will be enough to recognize the whole sentence.

From equation 5.8 we obtain

$$j = \log(p_s) / \log(p_w) \tag{5.9}$$

which can be used to calculate *j* from intelligibility scores.

Wagener et al. (1999c) found for the five-word German Oldenburger sentences that $j = 4.29$ for a signal-to-noise ratio of -5 dB SNR (recognition 80.7 %) and that j=3.18 for a signal-to-noise ratio -9 dB SNR (recognition 21.7 %). The authors also presented data for the Swedish Hagerman sentences where Hagerman found that $j = 4.77$ for recognition scores greater than 82 %, and that $j = 2.92$ for recognition scores lower than 28 %.

Based on the results for five-word sentences in noise described in sections 2.3.2 (first field test) and 2.3.4 (second laboratory test) and for three-word utterances in noise (section 3.3.3 from the second laboratory test), factor *j* could be calculated by pooling the results from all the subjects for each signal-to-noise ratio and is shown in Figure 5.2 together with the data from Wagener et al. and Hagerman.

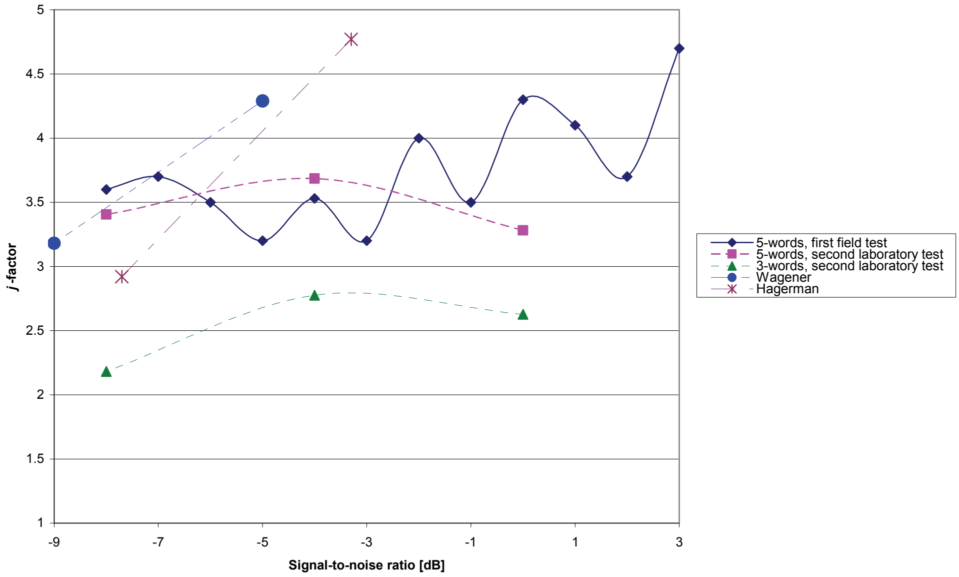


Figure 5.2 The *j*-factor calculated from listening tests in noise.

All the data points for the Norwegian material are based on testing with one list of 10 five-word sentences or 10 three-word utterances. The first field test was performed between 17 and 33 subjects except for the measurements for the condition -4 dB SNR which involved 106 subjects. The first field test was not performed with the final version of the material, but with the version prior to the level adjustments of the words described in section 2.2.2.5. The second laboratory test was performed on nine subjects.

In Figures 5.3-5.5 we apply the method used by Boothroyd and Nittrouer for zero predictability, low predictability and high predictability sentences to this data. For each list of sentences each subject's sentence probability(i.e., the estimated sentence probability from the sentence score) is presented as a function of the word probability (i.e., the estimated word probability from the word score) in a scatter diagram. The *j*-factor from all measurements with $0.1 \leq p_s \leq 0.9$ was calculated by equation 5.9, and the solid line show the prediction of equation 5.9 where *j* has been substituted by the mean value.

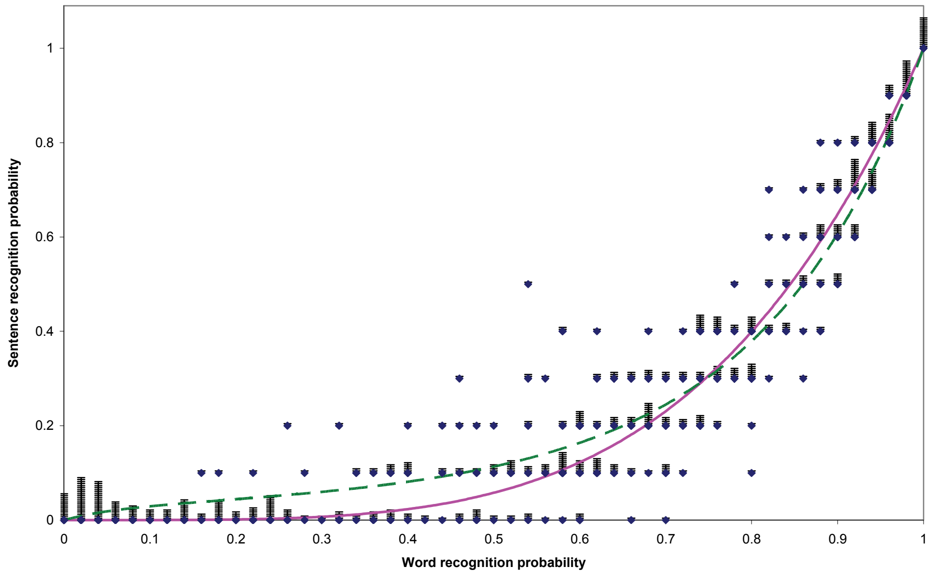


Figure 5.3 The relationship between recognition probabilities for words and sentences for five-word sentences in the first field test. Horizontal lines above each data point show the number of multiple identical data points. The fitted lines are $p_s = (p_w)^j$ where $j = 4.11$ (solid) or $j = 1.13 + 4.02 \cdot p_w$ (dashed).

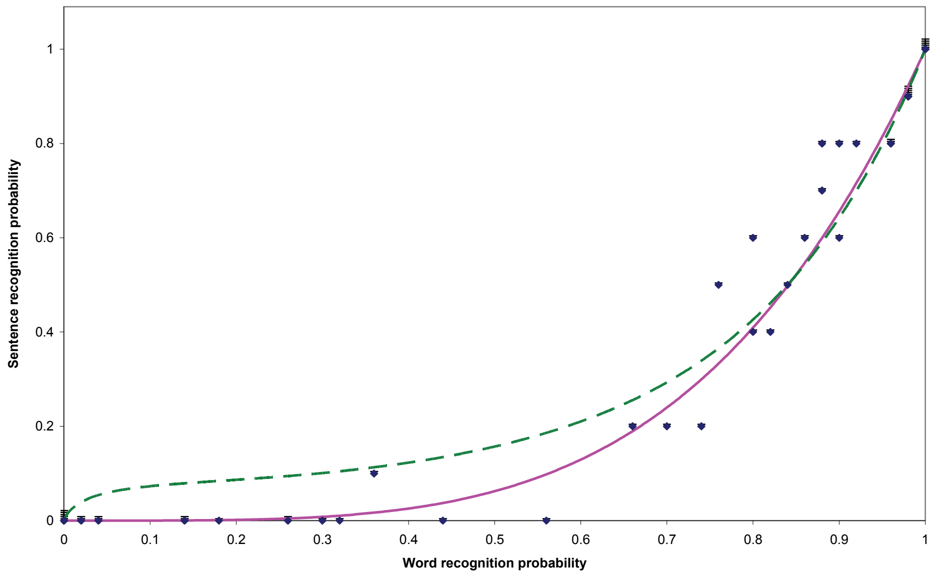


Figure 5.4 The relationship between recognition probabilities for words and sentences for five-word sentences in the second laboratory test. Horizontal lines above each data point show the number of multiple identical data points. The fitted lines are $p_s = (p_w)^j$ where $j = 4.00$ (solid) or $j = 0.75 + 3.84 \cdot p_w$ (dashed).

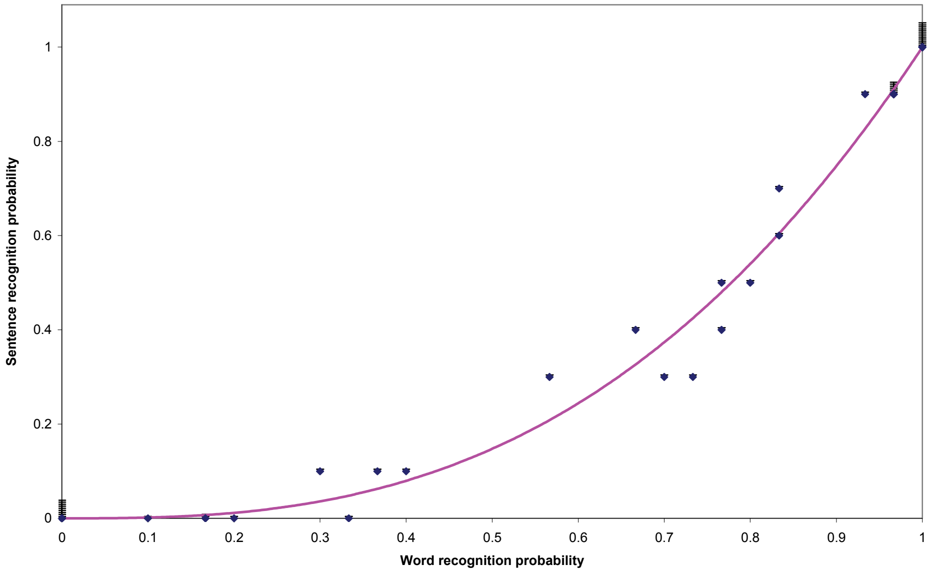


Figure 5.5 The relationship between recognition probabilities for words and sentences for three-word utterances in the second laboratory test. Horizontal lines above each data point show the number of multiple identical data points. The fitted line is $p_s= (p_w)^j$ where $j = 2.76$.

The results from the first field test depicted in Figure 5.3 are based on measurements performed on 106 subjects. A total number of 530 tests involving 10 five-word sentence lists was conducted with signal-to-noise ratios varying from -12 to +3 dB SNR. 340 of these tests satisfied the criteria that sentence scores should not be less than 10 % or greater than 90 % and j values should be calculable. A significant correlation was found between the value of j and the recognition probability of the words (r[338] = 0.546, $p < 0.001$) was found. j as a function of word recognition probability was found with linear regression as $j = 1.13 + 4.02 \cdot p_w$. Substituting this representation in equation 5.8 gives the predicted relationship between sentences and word recognition probability, which is shown with a dashed line in Figure 5.3. An alternative prediction of the relationship is shown with the solid line in Figure 5.3, where the mean value of j is inserted into equation 5.8. The mean value of j is 4.11, with 95 % confidence limits of ±0.14. The solid line seems to fit the scatter points better than the dashed line. The coefficient of correlation between measured sentence scores and those predicted from word scores is 0.9714 when the j from the linear regression is used, and 0.9709 when the mean value of j is used. The standard deviation of the difference between predicted and measured sentence recognition probability is 7.4 and 7.5 percentage points respectively. The conclusion is that both $j = 1.13 + 4.02 \cdot p_w$ and $j = 4.11$

could be used to predict sentence scores based on word scores in equation 5.8 with almost the same accuracy.

The results from the five-word sentences of the second laboratory test in Figure 5.4 are based on nine subjects. A total number of 45 tests involving 10 five-word sentence lists were conducted with signal-to-noise ratios of +10, 0 , -4, -8 and -12 dB SNR. 22 of these tests satisfied the criteria that sentence scores should not be less than 10 % or greater than 90 % and j values should be calculable. Here too, we found a significant correlation between the value of j and of the recognition probability of the words (r[20] = 0.431, $p <$ 0.05) was found. j as a function of word recognition probability was found with linear regression as $j = 0.75 + 3.84 \cdot p_w$. Substituting this representation into equation 5.8 gives the predicted relationship between sentences and word recognition probability, which is shown with a dashed line in Figure 5.4. The mean value of j is 4.00, with 95 % confidence limits of ±0.54. The solid line in Figure 5.4 shows the predicted relationship by substituting this value into equation 5.8, which seems to fit the scatter point better than the first prediction. The coefficient of correlation between measured sentence scores and those predicted from word scores is 0.982 when the j from the linear regression is used, and 0.987 when the mean value of j is used. The standard deviation of the difference between predicted and measured sentence recognition probability is 8.1 and 6.5 percentage points respectively. The conclusion is that using the mean value of $j = 4.00$ in equation 5.8 gives the best prediction of sentence scores based on word scores.

The results from the three-word utterances of the second laboratory test in Figure 5.5 are based on nine subjects. A total number of 45 tests involving 10 three-word utterance lists were conducted with signal-to-noise ratios of +10, 0 , -4, -8 and -12 dB SNR. 19 of these tests satisfied the criteria that sentence scores should not be less than 10 % or greater than 90 % and j values should be calculable. The coefficient of correlation between the value of j and of the word recognition probability failed to reach the 5 % level of significance (r[17] = 0.394); therefore there was no need to use linear regression here. The mean value of j is 2.76, with 95 % confidence limits of ±0.28. The solid line in Figure 5.5 shows the predicted relationship by substituting this value into equation 5.8. The coefficient of correlation between measured sentence scores and those predicted on the basis of word scores is 0.996 when the mean value of j is used. The standard deviation of the difference between predicted and measured sentence recognition probability is 3.9 percentage points. The conclusion is that using the mean value of $j = 2.76$ in equation 5.8 gives a very good prediction of sentence scores based on word scores.

Figures 5.2-5.5 present estimations of the j-factor determined by researchers from different institutions and for different languages, and for the Norwegian material in addition calculated by different methods. When trying to draw conclusions from the whole of this data we have to consider a remark made by Boothroyd and Nittrouer that the calculation of the *j*-factor is based on the estimation of recognition probability, which has a high test-retest variability unless obtained with a very large number of items. And the *j*-factor is essentially a difference score and therefore has an even greater test-retest variability. Each signal-to-noise ratio result in the tests conducted with the Norwegian material is based on that each test subject listened to only one test list consisting of 10 sentences. Although the procedure will cause some variability, this will be compensated for by the large number of subjects (106). The cyclical behaviour of the results from the first field test in Figure 5.2 can be a result of the variability in the method used to obtain the data. It can also be caused by the fact that all measurements for a specific signal-to-noise ratio were performed with the same list, and that the preliminary material was used before level adjustments had been made. Most of the results from the measurements using Norwegian five-word sentences lie between the lower and upper points for the Swedish and German material, but the slope of j for increasing signal-to-noise ratios is much lower for the Norwegian material. In fact, it seems that the j-factor is constant for the Norwegian material except for the best signal-to-noise ratios. The other way of calculating the j-factor done in connection with Figures 5.3-5.5 confirms this constant relationship. If we use the results for the Norwegian material in Figure 5.2 we would expect j ≈ 3.5 for the five-word sentences and j ≈ 2.5 for the three-word utterances. On the other hand, if we rely mainly on the second analysis we could stretch these values so that for the five-word sentences j ≈ 4 and for the three-word utterances j ≈ 2.8. This analysis shows that the five-word sentences and three-word utterances do not consist of, respectively, 5 and 3 independent test items, but that in practice they represent 3.5-4 and 2.5-2.8 items, respectively. The relative reduction in independent test items is greater for the five-word sentences (-20 % from 5 to 4 items) than for the three-word utterances (-7 % from 3 to 2.8 items).

In this chapter we are performing simulations of speech audiometry measurements involving both the five-word sentences and the three-word utterances based on binomial statistics. Based on the conclusion concerning the j-factor in the preceding paragraph it might seem correct to reduce the number of test items in a five-word sentence from 5 to 4 items. The decision not to reduce the number of test items was made on the basis of the following principles: When testing a sentence containing five words we can only obtain the responses of 0, 20, 40, 60, 80 or 100 % of the words correctly identified. If we should opt for 4 items instead of 5 for conducting the simulations we would obtain other scores which would give different behaviour when performing the simulations compared to the real test situations. But we have to keep in mind that the simulations are based on the principles that all test items have equal probability and that they do not influence each other, which are prerequisites that have been shown to be counterfactual. Hagerman (1976 and 1989) presents reasons for actually increasing the number of test items when performing simulations on speech discrimination: He finds that the variance is reduced by about 25 % when measuring speech discrimination scores compared to simulation predicted by binomial sampling. He therefore increases the number of test items in the simulations from 25 to 33 or from 50 to 66.

5.3 Methods

In section 5.4 results from simulated speech audiometry procedures performed on the four hypothetical subjects HS1-HS4 are presented. Three different methods for estimating the thresholds are evaluated, and for one of the methods the parameters for the fitted logistic functions are also estimated.

All simulations are based on the principle that when we are testing a word list at a specified level where all the words have the same probability of recognition, the distribution of correct responses will follow the binomial distribution. The procedure for simulating the response can be described by the following steps:

1. The probability of recognition for the words tested at the specified level is given by the logistic function. The probability is computed as the speech intelligibility, *SI(L)* by entering the level (*L*) and the parameters for the hypothetical subject (Table 5.1) into equations (5.2)-(5.4).
2. The probability mass functions (*f(k))* are computed by equation (5.5). $p = SI(L)/100$ is used as the probability, *n* is number of words in the list. The functions are computed for all possible values for *k*, where $0 \leq k \leq n$.
3. A random number with a rectangular distribution between 0 and 1 is generated and used to lookup one of the possible values *k* based on the probability mass functions computed in point 2. The *k* selected is the simulated response, and the statistics of repeated simulated responses will follow the binomial distribution.

This procedure for simulating a response when a list of speech audiometry words are presented at a specified level was programmed as a Matlab routine and is used as a building block for all the simulations presented in the following sections.

For many of the simulation results presented in section 5.4 the testing parameters, such as the number of levels tested or the number of words tested at each level, were adjusted. The adjustments were made in order to identify the least number of levels or words needed to acquire results with a defined accuracy for the speech recognition threshold. The defined accuracy for 95 % of the simulations was pragmatically chosen to lie within ±7.5 dB of the correct threshold which for our simulations was 35 dB.

5.3.1 ISO 8253-3 Determination of speech recognition threshold level, procedure A

This procedure (here called procedure A) is one of two proposed procedures in the ISO 8253-3 (1996) standard Acoustics - Audiometric test methods - Part 3: Speech audiometry. The procedure is described as a descending procedure using 5 dB steps. The other procedure (an alternative descending procedure using 2 dB or 5 dB steps) is not evaluated here.

Procedure A can be described by the following steps:

1. Familiarize the test subject with the procedure and material by presenting a number of test items at a sufficiently high level to be clearly audible, i.e. using a hearing level of speech of 20 to 30 dB above the average of the subject’s pure tone hearing threshold levels at 500, 1000 and 2000 Hz is proposed.
2. Reduce the speech levels in steps of 5 dB and present at least two test items on each level. Continue until you find the level where the test subject no longer responds correctly to all test items.
3. Present a set of test items at the level found in point 2 and record the number of correct responses. A set of test items must contain at least 10 items.
4. If the score is at least 50 %, reduce the level in steps of 5 dB and present a new set of test items until the score drops below 50 %. Usually one level is found to give scores of somewhat more than 50 % and the next level down will give scores of somewhat less than 50 %.
5. The speech recognition threshold can be computed by linear interpolation of the lowest level giving a score over 50 % and the highest level giving a score below 50 %.

A routine was developed in Matlab to simulate speech audiometry performed according to this procedure. Instead of testing on real subjects we conducted simulated tests on the hypothetical subjects defined by the parameters in Table 5.1. The responses were computed by the Matlab routine described in section 5.3. For each hypothetical subject HS1-HS4, 5000 simulations of speech audiometry procedures were performed. Statistics presenting the results are shown graphically in section 5.4.1. The selected starting point for our simulations was 60 dB, with random variations of ±5 dB.

This procedure estimates only the speech recognition threshold defined as the 50 % score. The correct estimates for the hypothetical subjects are HS1-HS2: 35 dB; and HS3-HS4: 37.6 dB.

5.3.2 The Hagerman and Kinnefors S/N-threshold method

Hagerman and Kinnefors (1995) developed the S/N-threshold method to speed up the measurement time required by Hagerman's material. The earlier recommendation was to use a full list of 10 five-word sentences in 3-dB steps to find the S/N threshold, defined as 50 % recognition. This new procedure aims at finding the threshold which is now defined as the 40 % recognition in the following way:

First one list (usually two to five sentences) was used in order to make adjustments to a comfortable level of speech. Then a training list containing ten sentences was presented. The signal-to-noise ratio was +20 dB for the first sentence, and was adjusted to +10, +5, 0, -5 and -8 dB for the succeeding sentences. After the sixth sentence, or earlier if only two words or fewer were recognized, an adaptive adjustment procedure was followed for the rest of the training list. The adaptive adjustment procedure was also used for the measurement list. Based on the number of words recognized in the preceding sentence, the change in signal-to-noise ratio for the succeeding sentence was given in Table 5.3. The signal-to-noise ratio was determined for ten sentences and the mean value was defined as the threshold.

Table 5.3 Rule for adjustment of signal-to-noise ratio after each sentence.

Number of correct words	0	1	2	3	4	5
Change of SNR, dB	+2	+1	0	-1	-2	-3

A routine was developed in Matlab to simulate speech audiometry performed according to this procedure. Instead of testing on real subjects we conducted simulated tests on the hypothetical subjects defined by the parameters in Table 5.1. The responses were computed by the Matlab routine described in section 5.3. For each hypothetical subject HS1-HS4, 2500 simulations of speech audiometry procedures were performed. Statistics presenting the results are shown graphically in section 5.4.2. The selected starting point for our simulations was 55 dB, with random variations of ±5 dB.

This procedure estimates only the speech recognition threshold defined as the 40 % score. The correct estimates for the hypothetical subjects are HS1: 34.0 dB; HS2: 31.6 dB; and HS3-HS4: 35.0 dB.

5.3.3 The Hagerman and Kinnefors SRT-threshold method

Hagerman and Kinnefors (1995) developed a method for measuring SRT in quiet based on principles similar to the ones described in the preceding section. The main difference is that Table 5.4 is used to calculate the changes in speech level, and the threshold was based on the mean value of only six sentences. The changes incorporated in Table 5.4 were made because the steepness of the performance-intensity function in silence was about half that of the function in noise.

Table 5.4 Rule for adjustment of speech level after each sentence.

Number of correct words	0	1	2	3	4	5
Change of speech level, dB	+4	+2	0	-2	-4	-6

A Matlab routine was developed in a way similar to the procedure described in the preceding section. The results are presented in section 5.4.3. The selected starting point for our simulations was 55 dB, with random variations of ±5 dB.

This procedure estimates only the speech recognition threshold defined as the 40 % score. The correct estimates for the hypothetical subjects are HS1: 34.0 dB; HS2: 31.6 dB; and HS3-HS4: 35.0 dB.

5.3.4 The Brand and Kollmeier A1 threshold method

Brand and Kollmeier (2002) developed the A1 threshold method as a generalization of the Hagerman and Kinnefors procedures described in the two preceding sections. This method is also an adaptive one, and the presentation levels are adjusted with decreasing steps as the results converges around the threshold – here defined as the signal-to-noise ratio giving a 50 % speech recognition score. After each presentation the change in level, *ΔL* is calculated by the following equation:

$$\Delta L = -\frac{f(i)\cdot(prev - tar)}{slope} \qquad (5.10)$$

where parameter *f(i)* is reduced for each reversal of the presentation level and parameter *i* is incremented for each reversal. *prev* is the score for the previous sentence (0-1), and *tar* is the target, which is here 0.5. The slope parameter was set to 0.15 dB^{-1} (15 %/dB). Brand and Kollmeier performed simulations and decided that *f(i)*=1.5·1.41^{-i} and limiting the final value of *f(i)* to 0.1 yielded optimal efficiency. The logistic function in equation 5.11 was chosen to represent the performance-intensity function:

$$p(L, L_{50}, s_{50}) = \frac{1}{1 + e^{4 \cdot s_{50} \cdot (L_{50} - L)}} \tag{5.11}$$

where L is the level, L_{50} is the 50 % speech recognition threshold and s_{50} is the slope at this threshold. This function was fit to measured or simulated data by a maximum-likelihood method. The likelihood of a performance-intensity function is:

$$l(p(L, L_{50}, s_{50})) = \prod_{k=1}^{m} p(L_k, L_{50}, s_{50})^{c(k)} \cdot [1 - p(L_k, L_{50}, s_{50})]^{1-c(k)} \tag{5.12}$$

where *c(k)*=1 if the word k is repeated correctly and *c(k)*=0 if not. Parameters L_{50} and s_{50} are varied to maximize $\log(l(p(L,L_{50},s_{50})))$, which gives the maximum likelihood discrimination function. The number of sentences used in this procedure could be adjusted in order to achieve measurements with the desired accuracy. Even if this procedure gives an estimate of the slope, the selection of test levels is made in such a way that the slope estimate is expected to be unreliable.

A routine was developed in Matlab to simulate speech audiometry performed according to this procedure on the hypothetical subjects defined by the parameters in Table 5.1. The responses were computed by the Matlab routine described in section 5.3. For each hypothetical subject HS1-HS4, 500 simulations of speech audiometry procedures were performed. Statistics presenting the results are shown graphically in section 5.4.4. The selected starting point for our simulations was randomized within the range 25-45 dB.

This procedure estimates both the speech recognition threshold defined as the 50 % score and the slope at the threshold. The correct estimates of the thresholds for the hypothetical subjects are HS1-HS2: 35.0 dB; and HS3-HS4: 37.6 dB. The correct estimates of the slopes are HS1: 10 %/dB; HS2: 3 %/dB; and HS3-HS4: 3.7 %/dB.

5.3.5 The Brand and Kollmeier A2 threshold and slope method

Brand and Kollmeier (2002) developed the A2 threshold and slope method to simultaneously measure both the threshold and the slope of the performance-intensity curve. The principles of the procedure are mostly the same as the ones described in the previous section. However instead of *tar* set to 0.5, two randomly interleaved independent tracks are measured with *tar* set to 0.2 and 0.8 in equation 5.10.

A Matlab routine was developed in a way similar to the procedure described in the preceding section. The results are presented in section 5.4.5. The selected starting point for our simulations was randomized within the range 25-45 dB.

This procedure estimates both the speech recognition threshold, defined as the 50 % score, and the slope at the threshold. The correct estimates of the thresholds for the hypothetical subjects are HS1-HS2: 35.0 dB; and HS3-HS4: 37.6 dB. The correct estimates of the slopes are HS1: 10 %/dB; HS2: 3 %/dB; and HS3-HS4: 3.7 %/dB.

5.3.6 Methods with a fixed number of items at test levels with constant intervals

Simulations were conducted for two different methods for selecting the test levels: a constant stimuli method and an adaptive stimuli method. For each of these two methods we calculated the results by two alternative routines, the curve fitting routine and the counting routine.

5.3.6.1 The constant stimuli method

Prior to testing, the decisions are made as to which test levels to use and how many words to test at each level. The routine presented in section 5.3 provides a score for each of the applied test levels. The procedure is then repeated 500 times to simulate 500 speech audiometry procedures with the same hypothetical subject. Section 5.3.6.3 and 5.3.6.4 describe the two alternative routines for calculating the results.

Each figure in section 5.4.6 presents the results of such a set of 500 simulations. At each level, 3, 5, 10, 30 or 50 words have been tested. The intervals between the test levels have been adjusted to obtain approximately the same accuracy for each of the test sets tested on our hypothetical subjects HS1-HS4.

In “HiST taleaudiometri” (Øygarden 2009) several so-called quick-speed tests based on the constant stimuli method are introduced. These tests are developed for both three-word utterances and five-word sentences. The quick-speed tracks on the audio CDs or the audio DVD contain a list of sentences/utterances where the level is reduced by a fixed number of dBs for each new sentence/utterance. In real measurement situations the counting routine will be used to calculate the results, but for the simulations the curve-fitting routine is also evaluated. The results are presented in the "HiST taleaudiometri" report (Øygarden 2009). A summary of the results is presented in section 5.4.7.

5.3.6.2 The adaptive stimuli method

The procedure we have chosen to called the adaptive stimuli method is by and large the method applied by Norwegian audiologists when measuring the performance-intensity curve. Usually the starting level is set to about 10 dB above the expected speech recognition threshold and 10 words are presented. If a score of 100 % is not obtained, the level is increased in 5 dB steps until the score reaches 100 %. The level is then reduced in 5 dB steps from the start level until the score approaches 0 %. A suitable sigmoid curve is then drawn with visual interpolation between the measured scores.

The high and low scores do not need to be 100 % and 0 %. For some of the tests in "HiST taleaudiometri" we have chosen to stop increasing the test levels when reaching scores above 80 %, and stop decreasing the test levels when the scores drop below 20 %. This can reduce the measuring time, and can also be beneficial in relation to test subjects with reduced maximum speech discrimination, i.e. when it is difficult to obtain a score of 100 %.

Matlab routines have been written to simulate such measurements. They are based on the routines described in section 5.3. The important difference is that with the adaptive stimuli procedure the levels are adjusted depending on the scores obtained, in contrast to the procedure in section 5.3.6.1, where all the levels were fixed prior to starting the test.

Simulations were performed for the hypothetical subjects HS1-HS4 for a set of 9-50 words used at each intensity level, with some variations of the level intervals. Both the curve-fitting routine and the counting routine were evaluated during these simulations. The results are presented in the "HiST taleaudiometri" report (Øygarden 2009). A summary of the results is presented in section 5.4.7.

5.3.6.3 The curve-fitting routine

If we were performing real-world speech audiometry and wanted to draw the performance-intensity function, we could test a list of items at several levels and draw a curve with visual interpolation between the measured scores. The curve-fitting routine has been developed to emulate the drawing of the performance-intensity function.

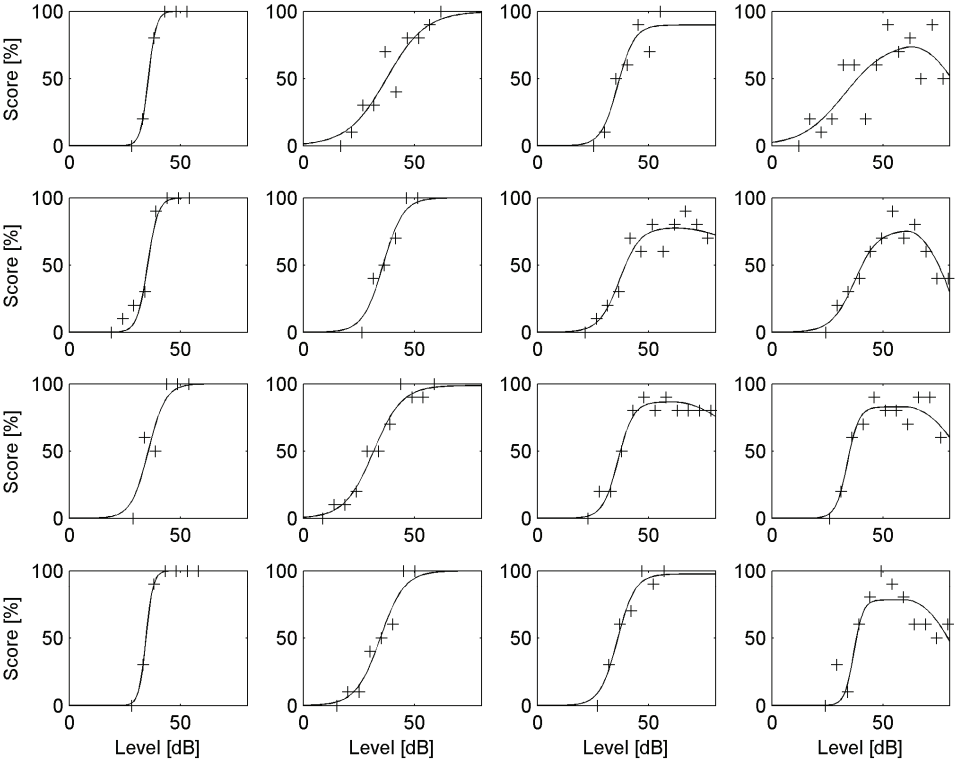


Figure 5.6 Four examples of logistic functions generated by the fitting procedure, (lines); and scores, (+) simulated for hypothetical subjects HS1-HS4 (in columns from left to right).

The constant stimuli method and the adaptive stimuli method described in the preceding sections simulate measurement methods which generate a set of scores obtained at stimulus levels with constant intervals. Matlab routines were developed to fit the logistic function given by equations (5.2)-(5.4) to the scores. The fitting procedure adjusts the parameters *slope*, L_{mid}, SI_{max} and *rof*. In order to search for the best possible fit of the logistic function, the procedures start with random values for *slope* and L_{mid}. The four parameters are adjusted one at a time to minimize the squared errors for the fitted curves by using the Matlab function fminbnd based on an algorithm called “Golden Section search and parabolic interpolation”. After having adjusted all the parameters once, the procedure is repeated, so that the parameters are adjusted once more. This adjustment procedure is repeated 15 times with different starting points for *slope* and L_{mid}. Then the solution giving the least square errors is selected as the result of fitting the logistic function to the set of simulated scores. Figure 5.6 presents some examples of logistic functions generated by this procedure for some simulated scores. The curves show a good fit with the scores, so this procedure seems to produce results which with reasonable accuracy reproduce how an audiologist could manually have drawn performance intensity curves from the measurement points.

The curve-fitting routine estimates all the parameters used to simulate the performance-intensity curves for our hypothetical subjects. The correct estimates of the thresholds for the hypothetical subjects are HS1-HS4: 35.0 dB. The correct estimates of the slopes are HS1: 10 %/dB, HS2: 3 %/dB and HS3-HS4: 5 %/dB. The correct estimates for the maximum speech recognition score are HS1-HS2: 100 %; and HS3-HS4: 80 %. The correct estimates for the rollover factor are HS1-HS3: 0; and HS4: 0.5.

5.3.6.4 The counting routine

When performing speech audiometry with a fixed number of items utilized at each level and using a constant step size between the levels, an estimate of the speech recognition threshold can be based on the total number of correct responses. In the present study this procedure will be called the counting routine.

If we start at level L_{start}, have tested n number of words at each level and decreased the level in ΔL steps, the speech recognition threshold can be estimated by the equation

$$L_{srt} = L_{start} + \frac{\Delta L}{2} - \frac{count \cdot \Delta L}{n} \qquad (5.13)$$

where *count* is the total number of correct responses.

The speech recognition thresholds estimated by the counting routine can be evaluated for all the simulations described in sections 5.3.6.1 and 5.3.6.2 since they are all based on test lists with a fixed number of words and tested with a constant step size between test levels. The counting routine is run parallel to the curve fitting routine in the same Matlab procedure.

The counting routine estimates only the speech recognition threshold defined as the 50 % score. The correct estimates for the hypothetical subjects are HS1-HS2: 35 dB; and HS3-HS4: 37.6 dB.

5.4 Results of simulations of speech audiometry measurements

Simulated speech audiometry was performed on all four hypothetical subjects HS1-HS4 with the different methods. Section 5.3 presented the correct threshold and other parameters for the simulations for each method.

5.4.1 ISO 8253-3 Determination of speech recognition threshold level, procedure A

The simulations were made using procedure A in the ISO standard for speech audiometry. The effects of some variations of the number of test items included in the set used at each level were explored.

5.4.1.1 10 items in test set

The results for the ISO 8253-3 Determination of speech recognition threshold level procedure A method are displayed in Figures 5.7-5.10. According to the standard the minimum number to be tested at each level is a set of 10 items.

As can be seen from Figure 5.7, only HS1 fulfils the requirement outlined in section 5.3 that 95 % of the thresholds must be located within ±7.5 dB of the correct 35 dB threshold, with lower and upper limits of 31.9- 37.9 dB.

The lower and upper limits for the other hypothetical subjects are 28.9-44.9 dB (HS2); 32.9-55.5 dB (HS3); and 32.9-54.7 dB (HS4). Especially the fact that the upper limits are found about 10-18 dB over the correct threshold shows that this procedure has severe limitations if the performance-intensity curve has a shallow slope (Figure 5.8), reduced maximum intelligibility (Figure 5.9) and/or rollover (Figure 5.10).

The mean number of tested items ranges from 38-48 for all the hypothetical subjects, but the spread is much larger for HS2, HS3-HS4 than for HS1.

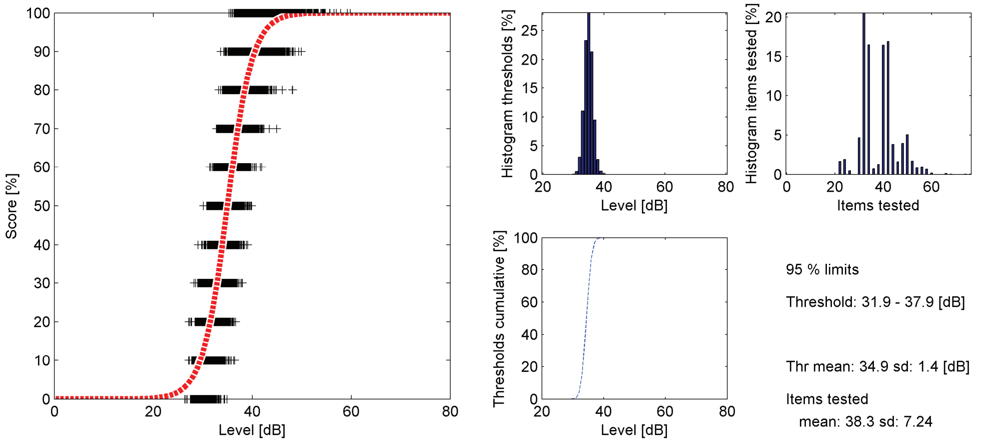


Figure 5.7 Simulated measurements of hypothetical subject HS1 according to ISO 8253-3 procedure A with 10 test items at each level. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 5000 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the number of items tested in each simulation. The 95 % limits for the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown.

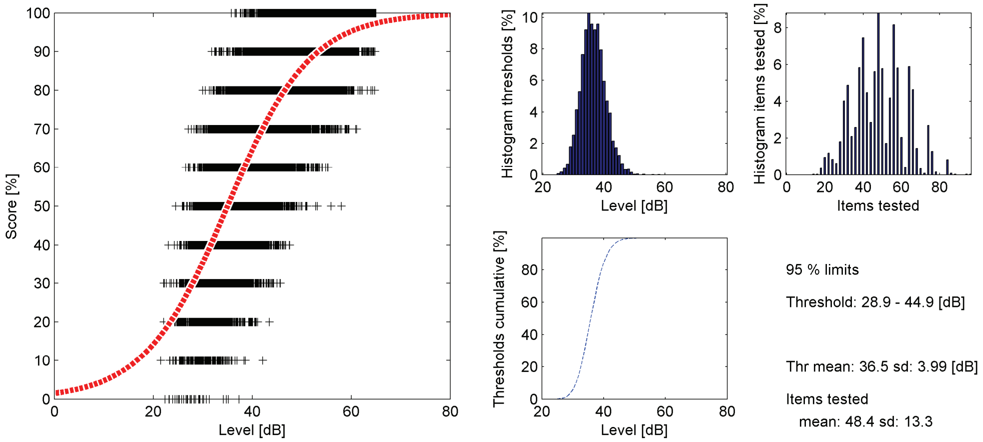


Figure 5.8 Results for HS2. Refer to Figure 5.7 for further explanation.

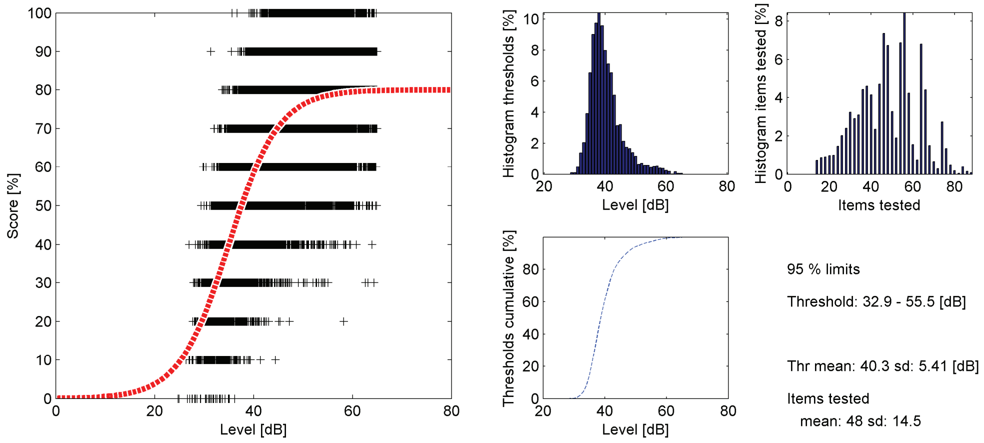


Figure 5.9 Results for HS3. Refer to Figure 5.7 for further explanation.

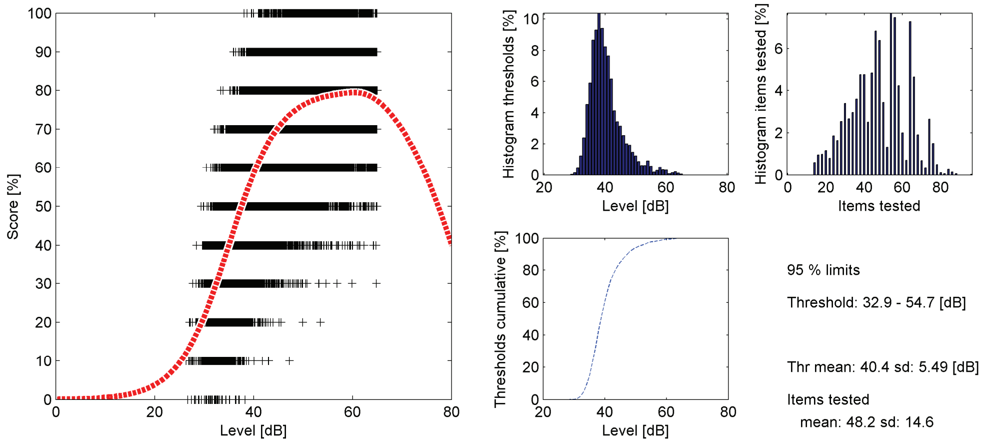


Figure 5.10 Results for HS4. Refer to Figure 5.7 for further explanation.

5.4.1.2 23 items in test set

The results for the ISO 8253-3 Determination of speech recognition threshold level procedure A method are displayed in Figures 5.11-5.14. The number of test items in the set used at each level had to be increased to 23 in order to get results where 95 % of the thresholds were found within ±7.5 dB of the correct threshold.

There is quite a substantial rise in the standard deviation of the threshold from HS1 (Figure 5.11) with a value below 1 dB to 2.44-2.66 for HS2-HS4 (Figure 5.12-5.14). This demonstrates that the method loses accuracy when the performance-intensity curve has a shallow slope, reduced maximum intelligibility and/or rollover. The method compensates to some degree for the more difficult measurements by increasing the mean number of tested items from 76 for HS1 to between 112-117 for HS2-HS4. There is also a great variability in the number of items used in each measurement. The number of items differs by roughly a factor of three (n ≈ 180 / n ≈ 60) for HS2-HS3.

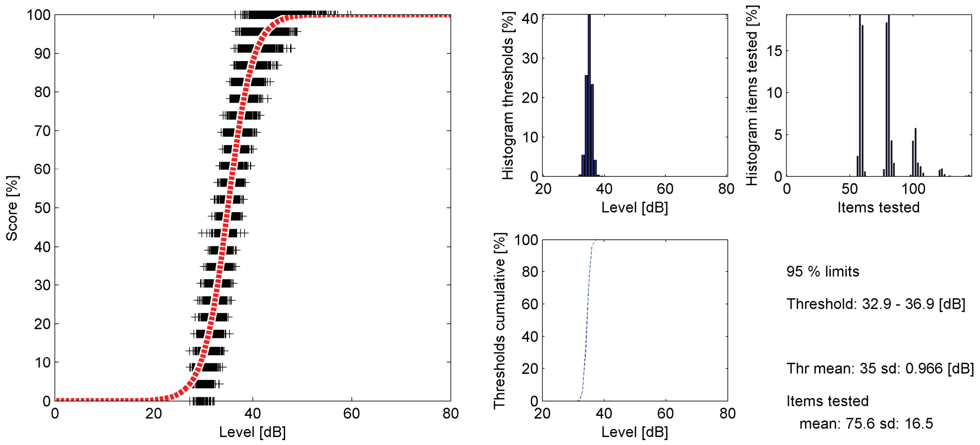


Figure 5.11 Simulated measurements of hypothetical subject HS1 according to ISO 8253-3 procedure A with 23 test items at each level. Refer to Figure 5.7 for further explanation.

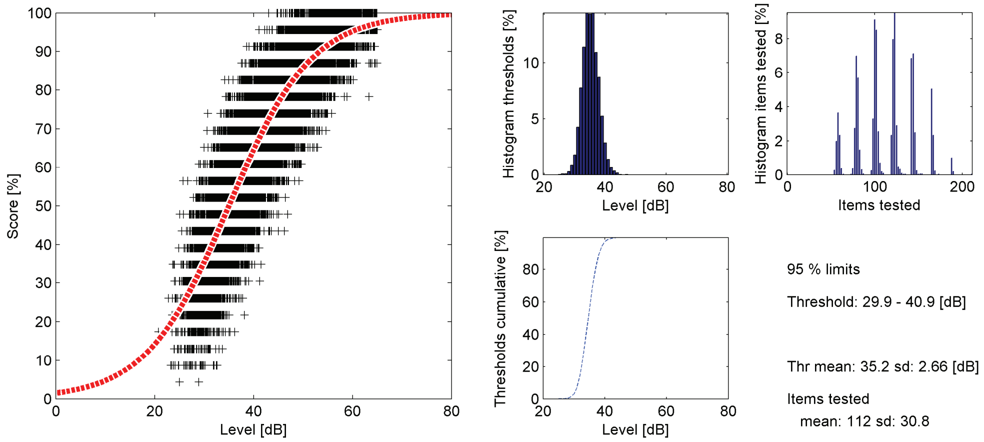


Figure 5.12 Results for HS2. Refer to Figure 5.11 for further explanation.

Simulations were also performed for the 16 %/dB slope as expected when we use three-word utterances in noise with 30 words in each set, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Thresholds: 95% 33.9-35.9 dB, mean 35 and standard deviation 0.665. Items tested: mean 93.5, standard deviation 16.4. Likewise, simulations were performed for the 14 %/dB slope as expected when we use five-word sentences in noise with 50 words in each set, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Thresholds: 95% 33.9-35.9 dB, mean 35 dB and standard deviation 0.576 dB. Items tested: mean 167, standard deviation 29.6.

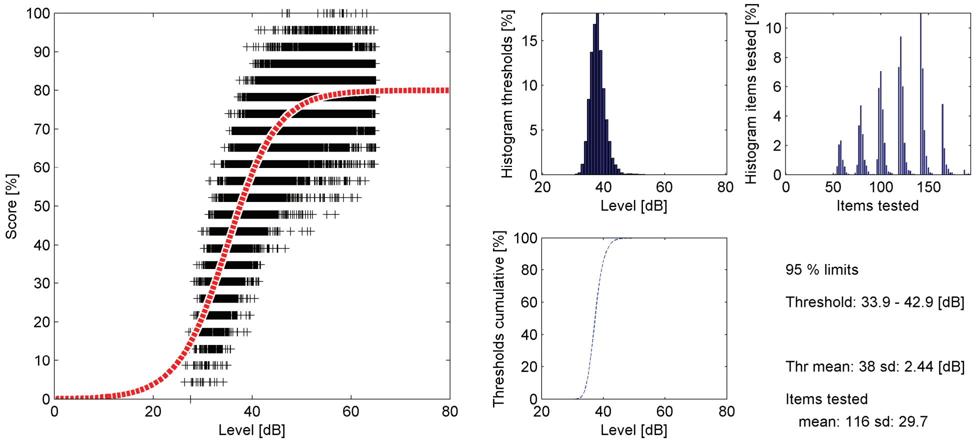


Figure 5.13 Results for HS3. Refer to Figure 5.11 for further explanation.

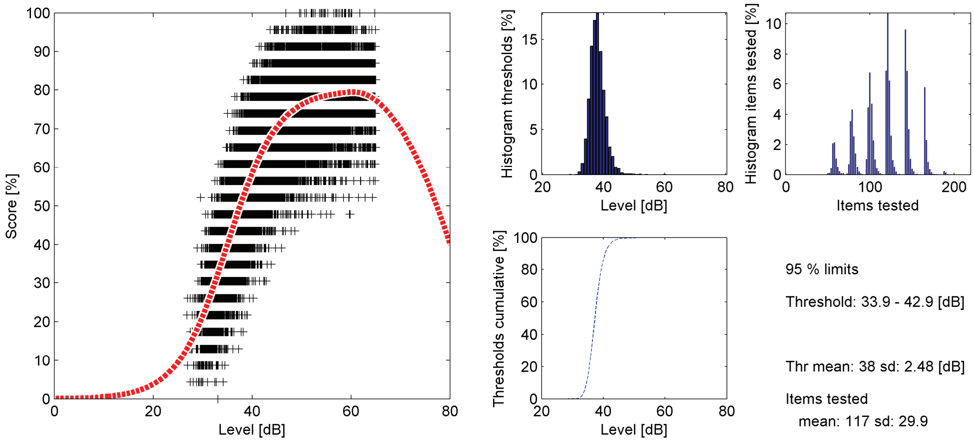


Figure 5.14 Results for HS4. Refer to Figure 5.11 for further explanation.

5.4.2 The Hagerman and Kinnefors S/N-threshold method

The results for the Hagerman and Kinnefors S/N-threshold method are displayed in Figures 5.15-5.18. It appeared that this procedure gave very good results for hypothetical subject HS1, with a standard deviation of only 0.74 dB. Hypothetical subject HS2-HS4 was within +5 dB and -4 dB of the correct threshold for 95 % of simulations, but the standard deviation

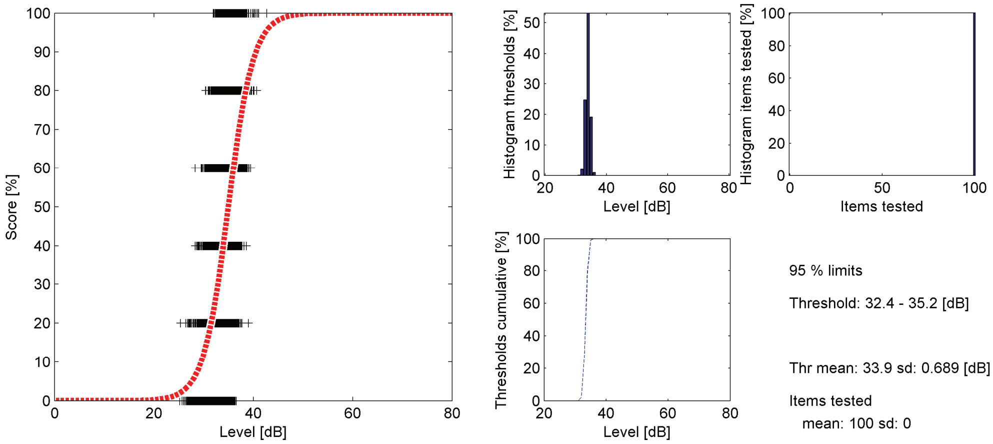


Figure 5.15 Simulated measurements of hypothetical subject HS1 according to the Hagerman and Kinnefors S/N-threshold method. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 2500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of number of items tested in each simulation. The 95 % limits for the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown.

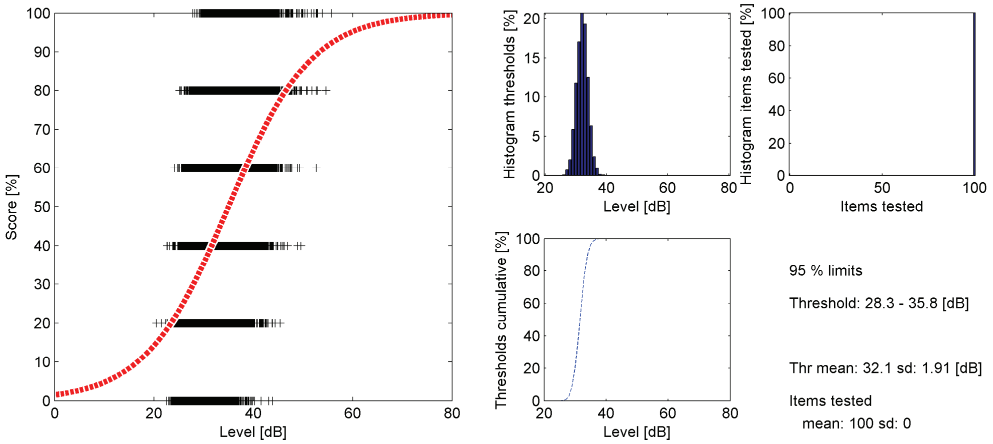


Figure 5.16 Results for HS2. Refer to Figure 5.15 for further explanation.

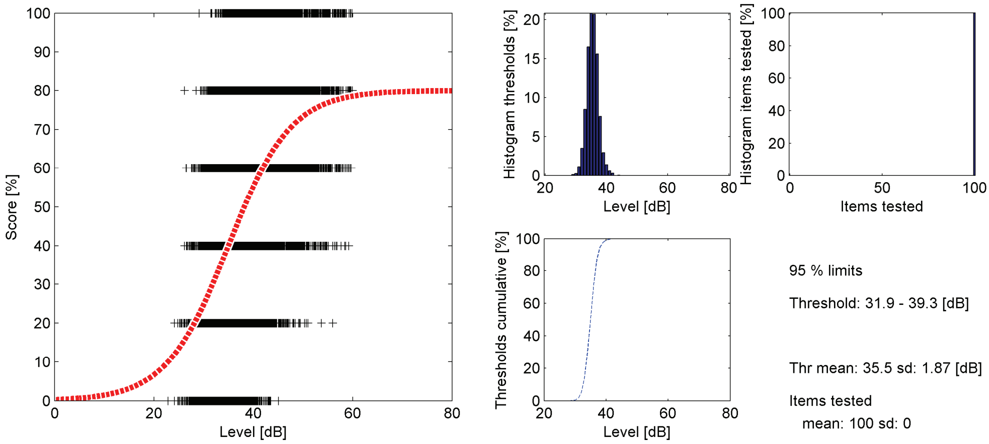


Figure 5.17 Results for HS3. Refer to Figure 5.15 for further explanation.

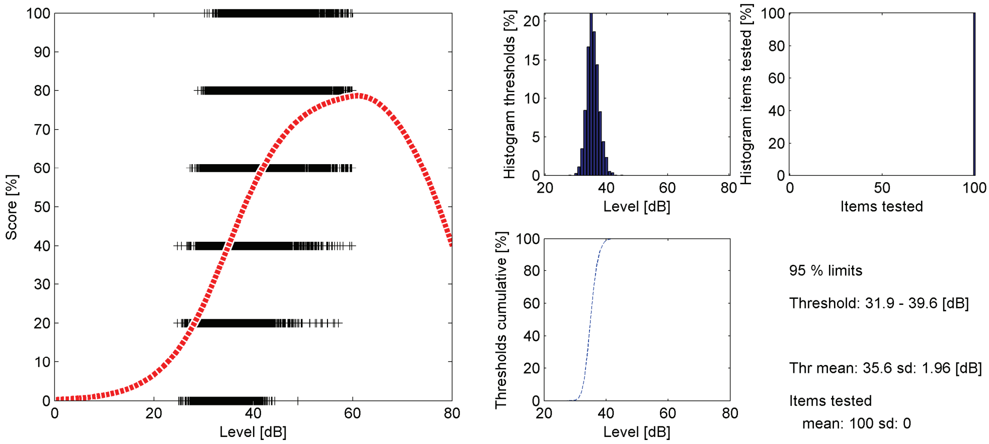


Figure 5.18 Results for HS4. Refer to Figure 5.15 for further explanation.

increased to 1.87-1.96 dB. This demonstrates that the method gives very accurate results, but the standard deviation increases when the performance-intensity curve has a low slope, is reduced in terms of maximum intelligibility and/or due to the effects of rollover. The number of tested items is 100 for all the measurements, which means that 20 five-word sentences are usually used.

Simulations were also performed for the 14 %/dB as expected with five-word sentences in noise, with the rest of the parameters kept as before for HS1. These simulations produced the following results. Thresholds: 95% 33.1-35.1 dB, mean 34.2 dB and standard deviation 0.513 dB. Items tested: 100.

5.4.3 The Hagerman and Kinnefors SRT-threshold method

The results for the Hagerman and Kinnefors SRT-threshold method are displayed in Figures 5.19-5.22. The results for this method are less accurate than the results of the related SN-method in the preceding section.

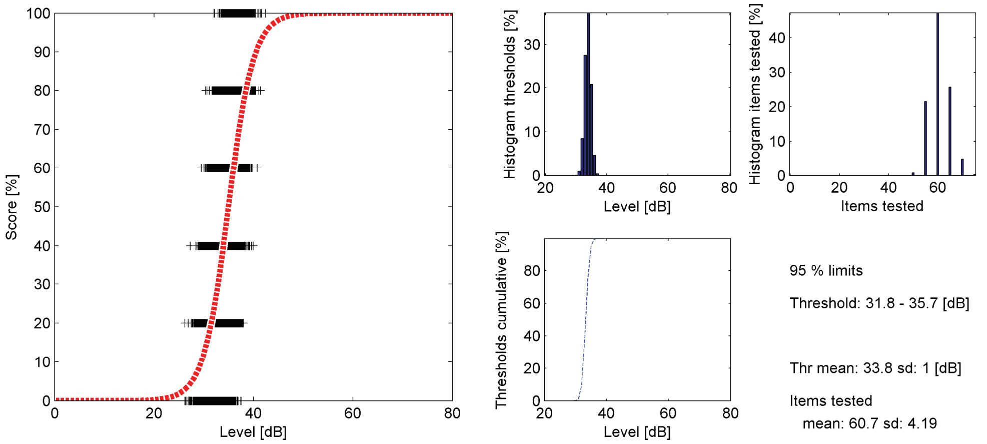


Figure 5.19 Simulated measurements of hypothetical subject HS1 according to the Hagerman and Kinnefors SRT-threshold method. The large panel shows the logistic function. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The middle top panel shows the histogram of the thresholds obtained during the 2500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of number of items tested in each simulation. The 95 % limits for the threshold plus mean and standard deviation for the threshold are indicated. Mean and standard deviation for the number of items tested are also shown.

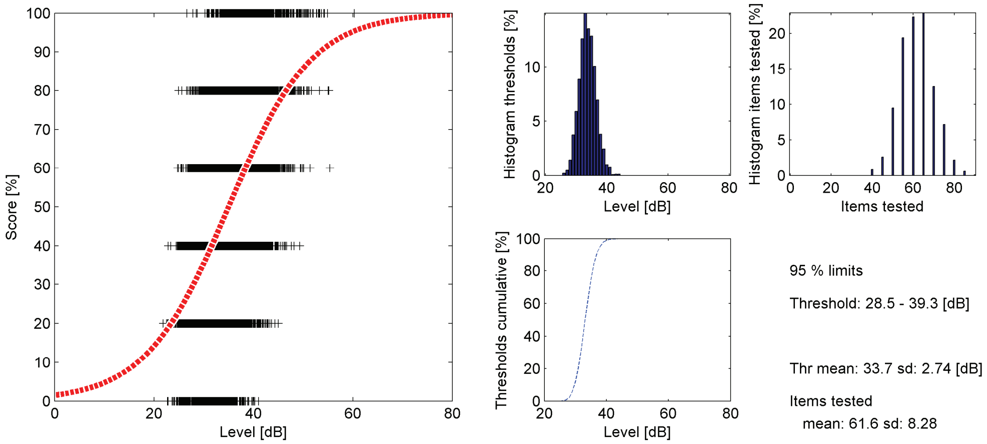


Figure 5.20 Results for HS2. Refer to Figure 5.19 for further explanation.

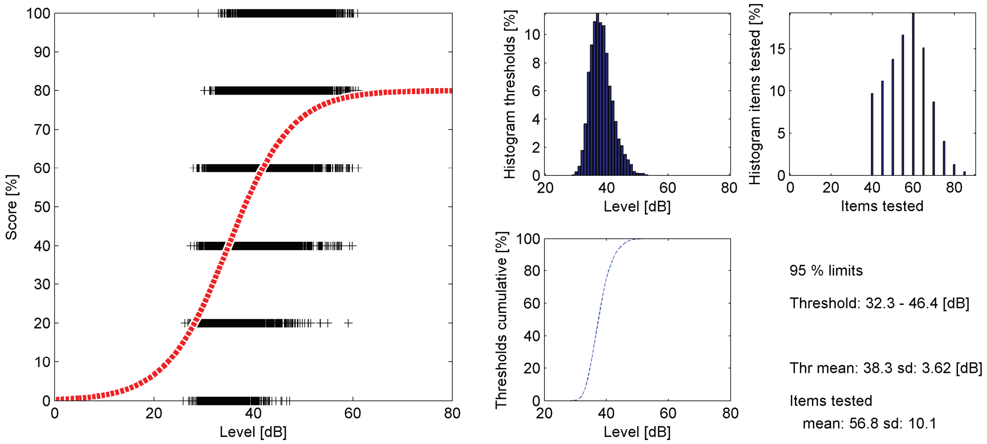


Figure 5.21 Results for HS3. Refer to Figure 5.19 for further explanation.

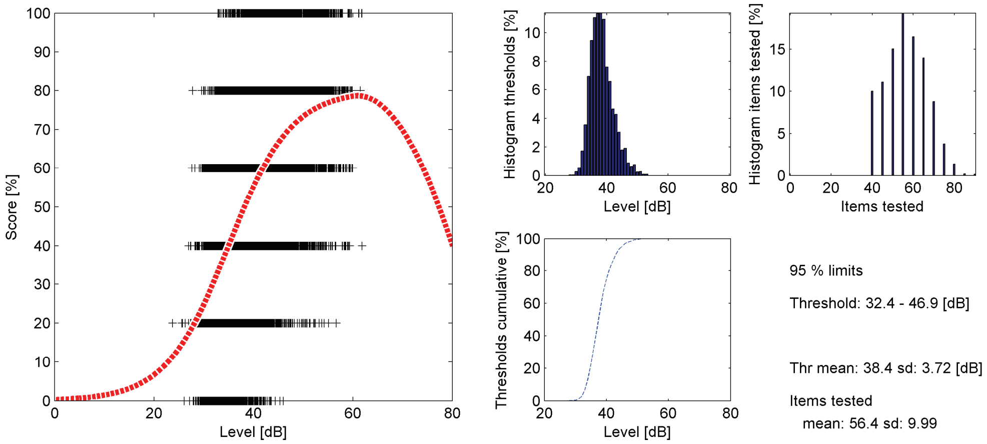


Figure 5.22 Results for HS4. Refer to Figure 5.19 for further explanation.

Hypothetical subject HS1 are within the 2.2 dB of correct result, but for HS2-HS4 the upper 95 % limit is 7.7-11.9 dB over the correct result. This demonstrates that the method loses accuracy when the performance-intensity curve has a low slope, is reduced in terms of maximum intelligibility and/or due to the effects of rollover. The mean number of tested items is around 60 for all the measurements, which means that 8-16 five-word sentences are usually used.

Simulations were also performed for the 14 %/dB slope as expected with five-word sentences in noise, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Thresholds: 95% 32.4-35.4 dB, mean 34.1 dB and standard deviation 0.773 dB. Items tested: mean 60, standard deviation 3.67.

5.4.4 The Brand and Kollmeier A1 threshold method

The results for the Brand and Kollmeier A1 threshold method are displayed in Figures 5.23-5.26. For this method it is possible to adjust the number of sentences used to achieve the desired accuracy. Here we used 19 sentences

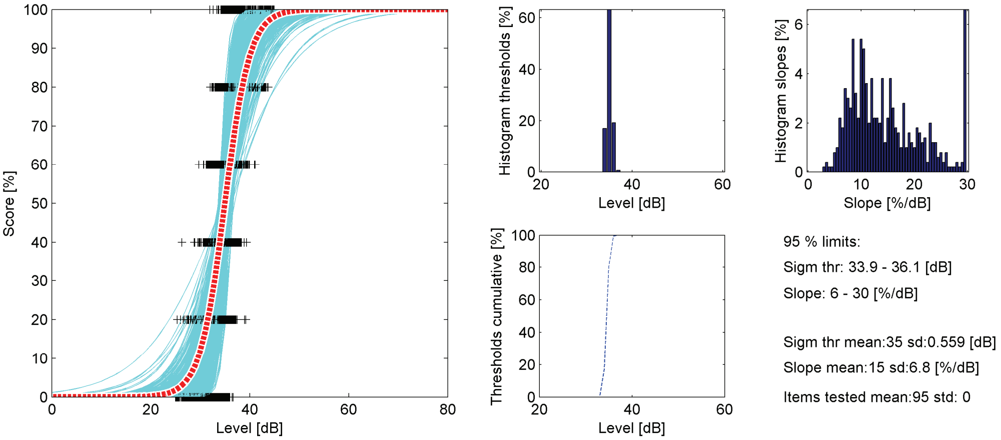


Figure 5.23 Simulated measurements of hypothetical subject HS1 according to the Brand and Kollmeier A1 threshold method, using 19 five-word sentences. The large panel shows the logistic function for the hypothetical subject as a thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The middle top panel shows the histogram of the thresholds obtained during the 500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the estimated slopes. The 95 % limits for the threshold and the slope plus mean and standard deviation for the threshold and the slope are indicated. Mean and standard deviation for the number of items tested are also shown.

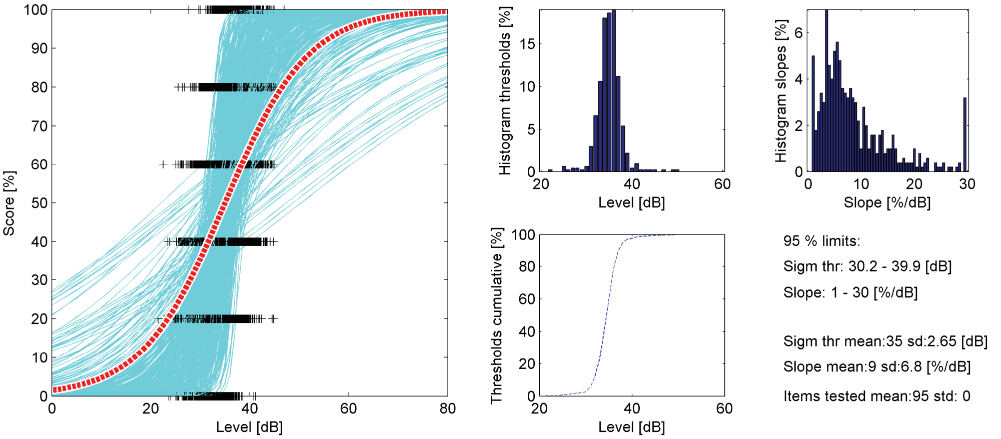


Figure 5.24 Results for HS2. Refer to Figure 5.23 for further explanation.

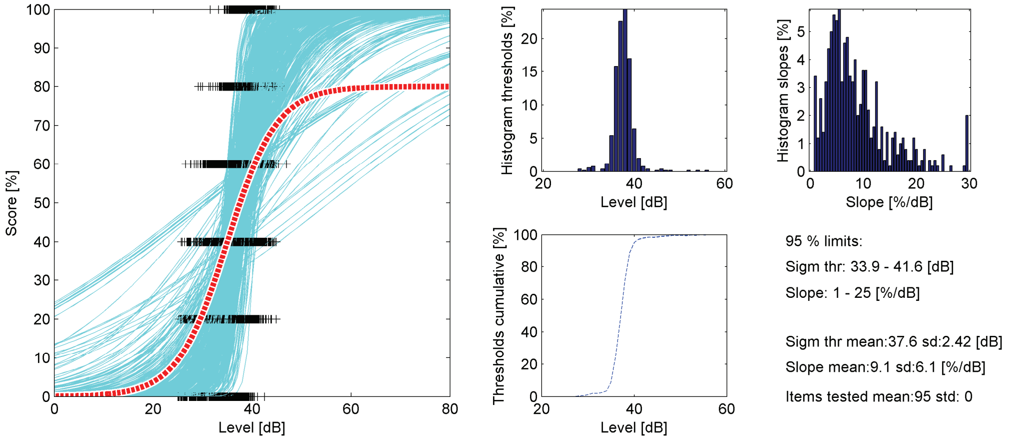


Figure 5.25 Results for HS3. Refer to Figure 5.23 for further explanation.

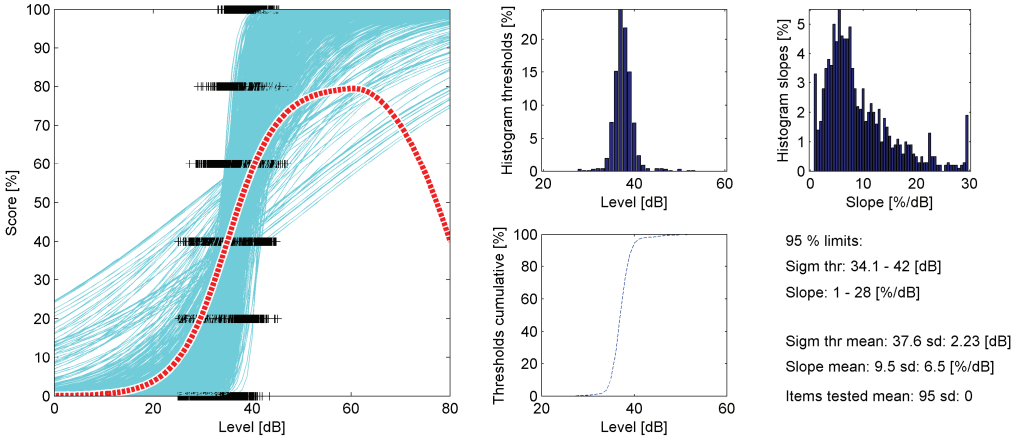


Figure 5.26 Results for HS4. Refer to Figure 5.23 for further explanation.

which was found to be just sufficient to locate 95 % of the simulations within the required ±7.5 dB range. Actually, almost all our simulations were located within ±5 dB of the correct result. The mean value of the thresholds was also correct for all four hypothetical subjects. The standard deviation was very good, with a value 0.56 dB for HS1 rising to about 2.2-2.7 dB for HS2-HS4. As expected, there were large variations for the estimated slopes with this method since the measurement levels were selected with the exclusive goal of locating the correct threshold.

Simulations were also performed for the 14 %/dB slope as expected with five-word sentences in noise, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Thresholds: 95% 34.3-35.9 dB, mean 35 dB and standard deviation 0.406 dB.

5.4.5 The Brand and Kollmeier A2 threshold and slope method

The results for the Brand and Kollmeier A1 threshold and slope method are displayed in Figures 5.27-5.30. For this method it is also possible to adjust

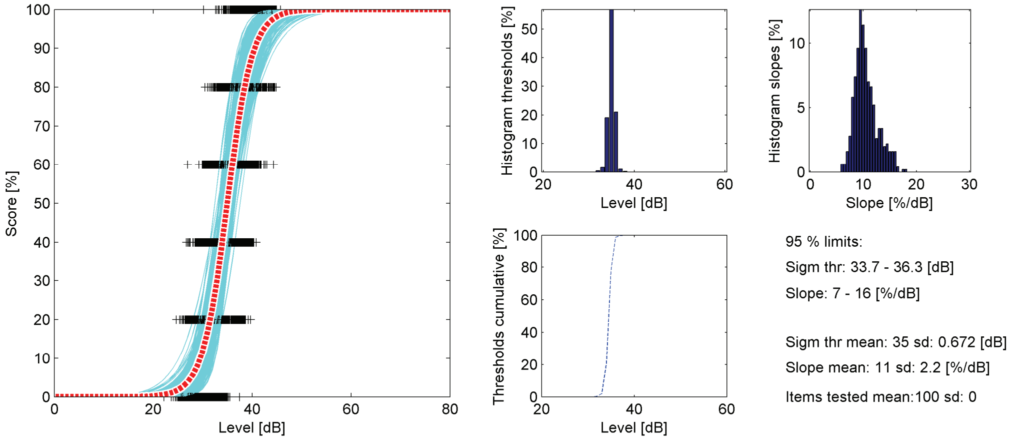


Figure 5.27 Simulated measurements of hypothetical subject HS1 according to the Brand and Kollmeier A2 threshold and slope method, using 20 five-word sentences. The large panel shows the logistic function for the hypothetical subject, indicated with a thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The middle top panel shows the histogram of the thresholds obtained during the 500 simulations, with the cumulative distribution of the thresholds in the panel below. The right panel shows the histogram of the estimated slopes. The 95 % limits for the threshold and the slope plus mean and standard deviation for the threshold and the slope are indicated. Mean and standard deviation for the number of items tested are also shown.

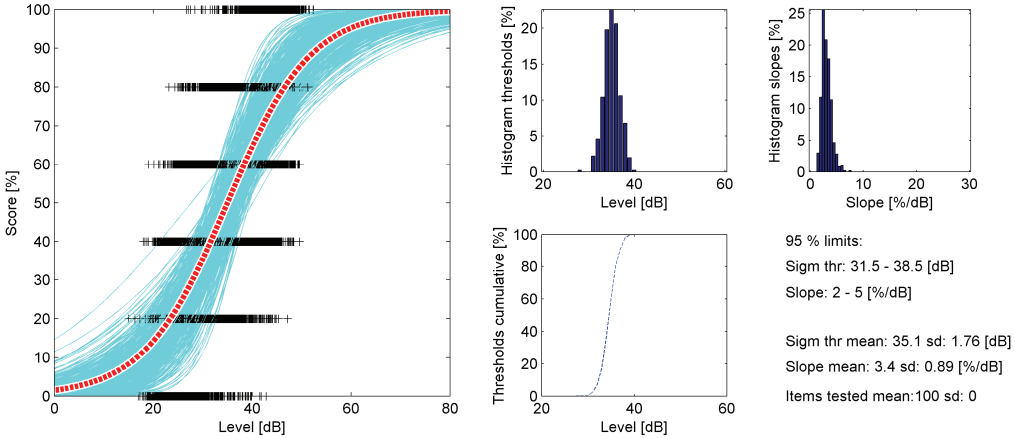


Figure 5.28 Results for HS2. Refer to Figure 5.27 for further explanation.

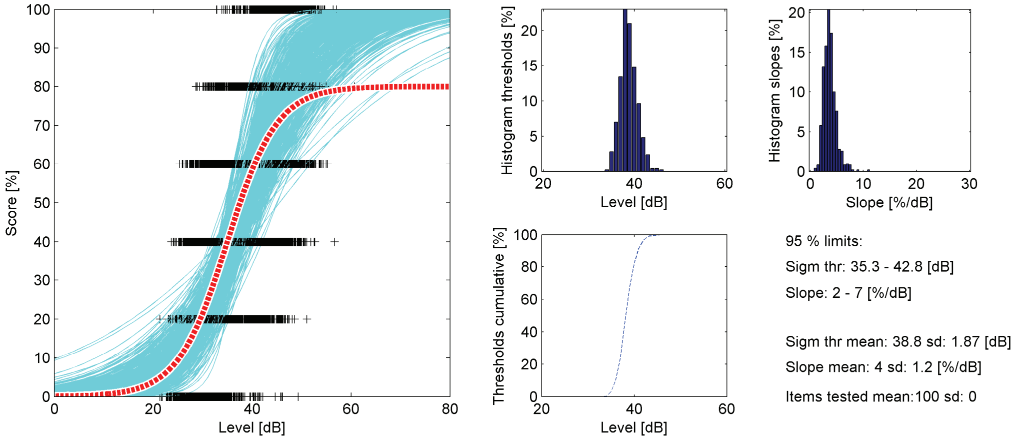


Figure 5.29 Results for HS3. Refer to Figure 5.27 for further explanation.

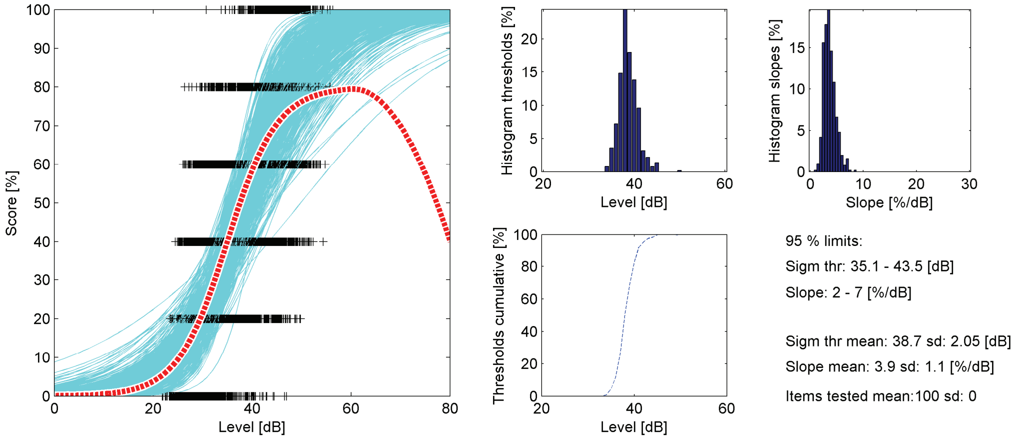


Figure 5.30 Results for HS4. Refer to Figure 5.27 for further explanation.

the number of sentences used to achieve the desired accuracy (cf. 5.4.4). Here were used 20 sentences, which yielded results only narrowly missing the desired ±7.5 dB range for 95 % of measurements: for the hypothetical subjects HS3-HS4 the results fell just outside this requirement. Increasing the number of sentences did not improve these results. The problem can be explained by the fact that HS3-HS4 have a maximum intelligibility of 80 %, while the method seeks to measure the 80 % point on the performance intensity curve.

The standard deviations for threshold varied between 0.67-2.1 dB, and the standard deviation values for the slopes were very good at between 0.9-2.2 %/dB.

Simulations were also performed for the 14 %/dB slope as expected with five-word sentences in noise, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Thresholds:

95% 34.1-36.0 dB, mean 35 dB and standard deviation 0.486 dB. Slopes: 95% 10-22 %/dB, mean 15 %/dB and standard deviation 2 %/dB.

5.4.6 The constant stimuli method

5.4.6.1 3 word sets at 1.5 dB intervals, 141 words/session

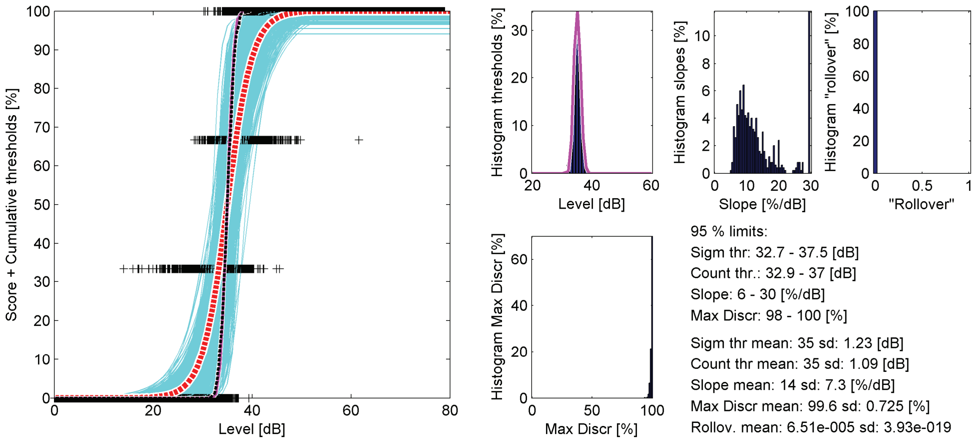
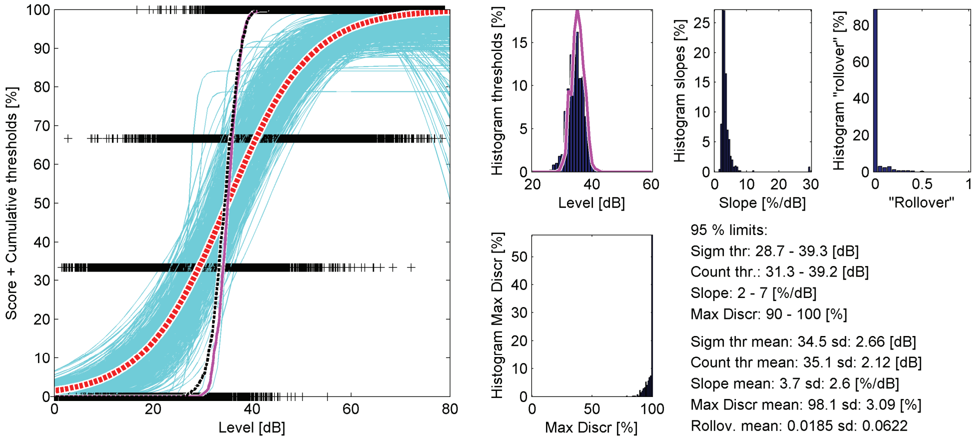


Figure 5.31 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 3 test items measured at every level from 5-75 dB in 1.5 dB intervals. The large panel shows the logistic function for the hypothetical subject, indicated by the thick dashed line. Plus signs indicate all the simulated scores obtained when "testing" at a specific level. Repeated identical scores cannot be discerned from a single score. The thin lines show the logistic curves fitted to the scores. The medium lines show the cumulative distribution of the thresholds estimated by the counting routine, solid line; and the curve-fitting routine, dashed line. The small top left panel shows the histogram of the thresholds obtained during the 500 simulations by the counting routine, solid line; and the curve-fitting routine, bar graph. The small top middle panel shows the histogram of the estimated slopes. The small top right panel shows the histogram of the estimated rollover parameter. The small bottom panel shows the histogram of the estimated maximum recognition score. The 95 % limits and/or means + standard deviations of the estimated parameters are indicated.

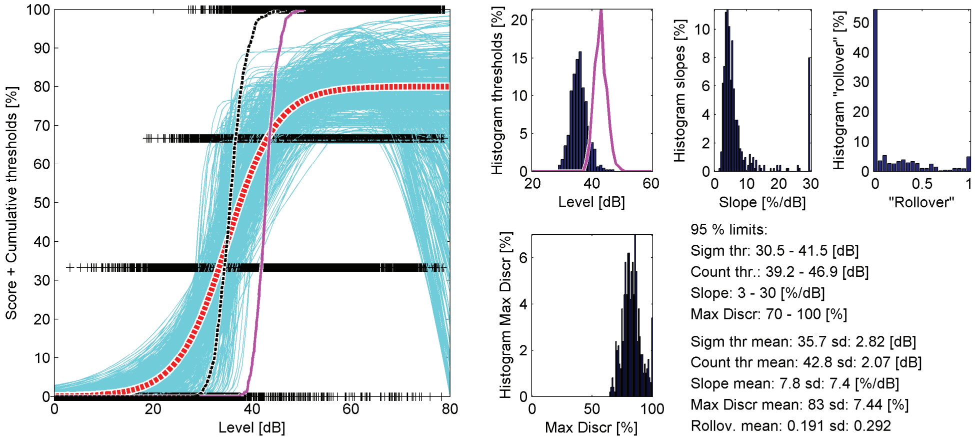
The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in Figures 5.31-5.34. In order to achieve results within the ±7.5 dB range when measuring between 5-75 dB for three-word utterances the interval between the levels had to be fixed at 1.5 dB. 141 items were used for each measurement. For the curve-fitting routine the standard deviation of the thresholds increases from 1.2 dB for HS1 to 2.7-2.9 dB for HS2-4. The counting routine produces lower standard deviations for the thresholds, but its mean values for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are not very accurate and the variability is large.



95 % limits:
Sigm thr: 28.7 - 39.3 [dB]
Count thr.: 31.3 - 39.2 [dB]
Slope: 2 - 7 [%/dB]
Max Discr: 90 - 100 [%]

Sigm thr mean: 34.5 sd: 2.66 [dB]
Count thr mean: 35.1 sd: 2.12 [dB]
Slope mean: 3.7 sd: 2.6 [%/dB]
Max Discr mean: 98.1 sd: 3.09 [%]
Rollov. mean: 0.0185 sd: 0.0622

Figure 5.32 Results for HS2. Refer to Figure 5.31 for further explanation.



95 % limits:
Sigm thr: 30.5 - 41.5 [dB]
Count thr.: 39.2 - 46.9 [dB]
Slope: 3 - 30 [%/dB]
Max Discr: 70 - 100 [%]

Sigm thr mean: 35.7 sd: 2.82 [dB]
Count thr mean: 42.8 sd: 2.07 [dB]
Slope mean: 7.8 sd: 7.4 [%/dB]
Max Discr mean: 83 sd: 7.44 [%]
Rollov. mean: 0.191 sd: 0.292

Figure 5.33 Results for HS3. Refer to Figure 5.31 for further explanation.

Simulations were also performed for the 16 %/dB slope as expected with three-word utterances in noise, with the rest of the parameters kept as before for HS1. This produced the following results: Sigmoid thresholds: 95 % 33.1-36.7 dB, mean 34.9 dB and standard deviation 0.945 dB. Count thresholds: 95% 33.3-36.4 dB, mean 34.9 dB and standard deviation 0.827 dB.

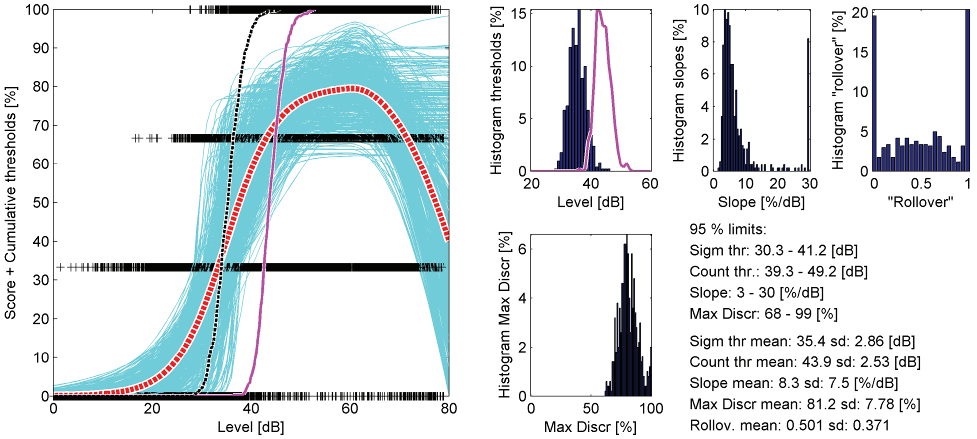


Figure 5.34 Results for HS4. Refer to Figure 5.31 for further explanation.

5.4.6.2 5-word sets at 2.5 dB intervals, 145 words/session

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in Figures 5.35-5.28. In order to achieve results within the ±7.5 dB when measuring between 5-75 dB for five-word sentences the interval between

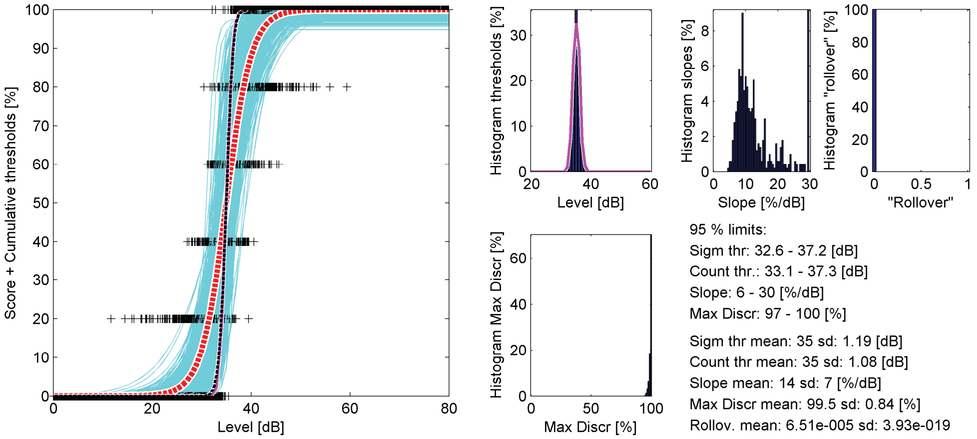


Figure 5.35 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 5 test items in each set were measured at every level from 5-75 dB in 2.5 dB intervals. Refer to Figure 5.31 for explanation the remaining details.

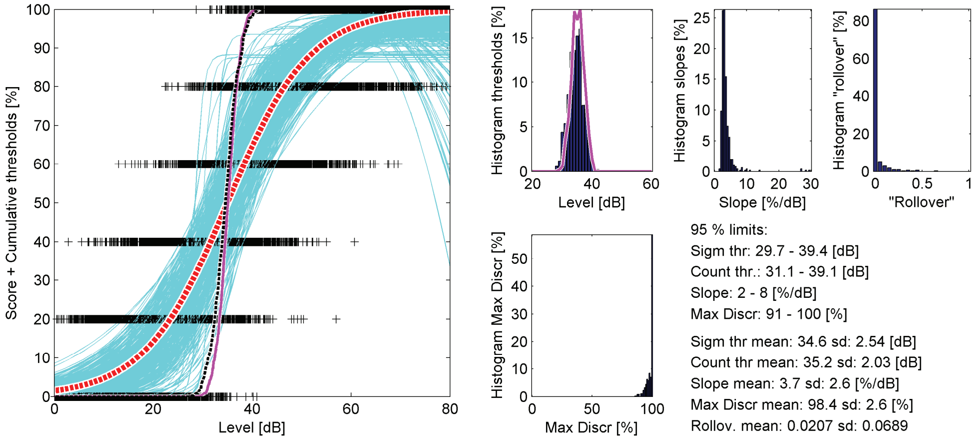


Figure 5.36 Results for HS2. Refer to Figure 5.35 for further explanation.

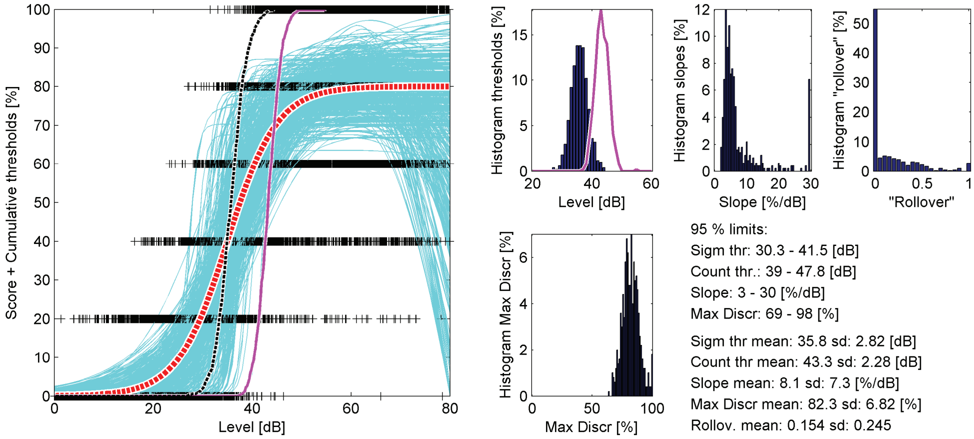


Figure 5.37 Results for HS3. Refer to Figure 5.35 for further explanation.

the levels had to be fixed at 2.5 dB. 145 items were used for each measurement. For the curve-fitting routine the standard deviation of the thresholds increases from 1.2 dB for HS1 to 2.5-2.9 dB for HS2-4. The counting routine results in lower standard deviations for the thresholds, but the mean values for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameter are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are not very accurate and the variability is large.

Simulations were also performed for the 14 %/dB slope as expected with five-word sentences in noise, with the rest of the parameters kept as before for HS1. These simulations produced the following results: Sigmoid thresholds: 95% 32.9-37.2 dB, mean 35 dB and standard deviation 1.1 dB.

Count thresholds: 95% 33.2-36.9 dB, mean 35 dB and standard deviation 0.988 dB.

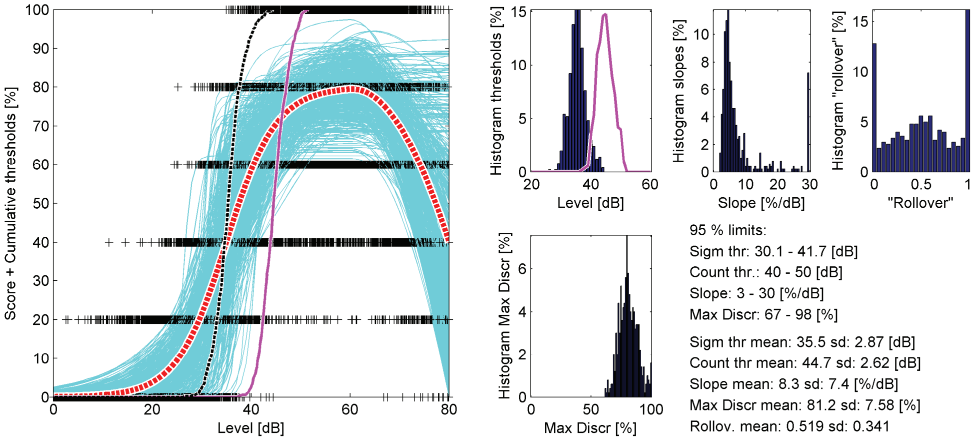


Figure 5.38 Results for HS4. Refer to Figure 5.35 for further explanation.

5.4.6.3 10 word sets at 7 dB intervals, 110 words/session

The results for the constant stimuli method with estimates calculated by both the curve fitting routine and the counting routine are displayed in Figures 5.39-5.42. In order to achieve results within the ±7.5 dB range when

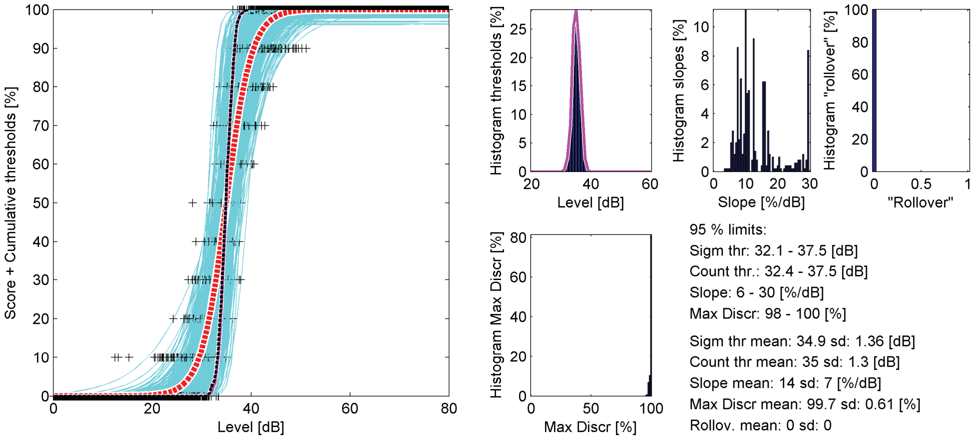


Figure 5.39 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve fitting routine and the counting routine. 10 test items in each set measured at every level from 5-75 dB in 7 dB intervals. Refer to Figure 5.31 for explanation of the remaining details.

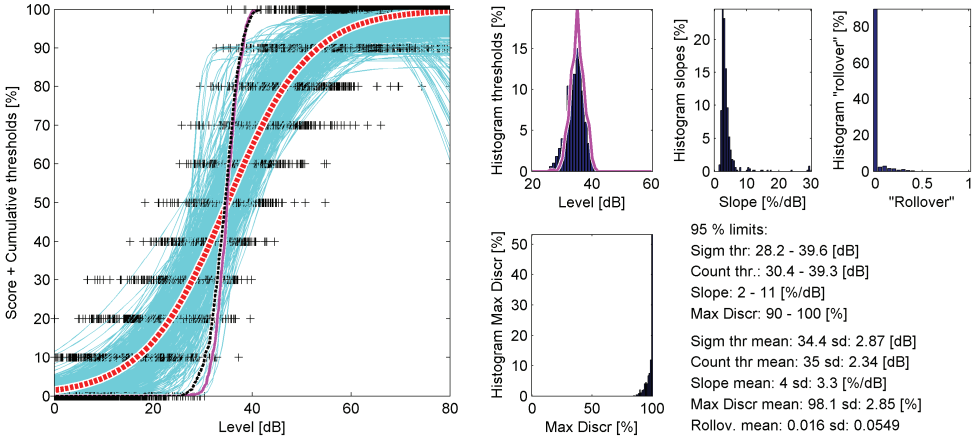


Figure 5.40 Results for HS2. Refer to Figure 5.39 for further explanation.

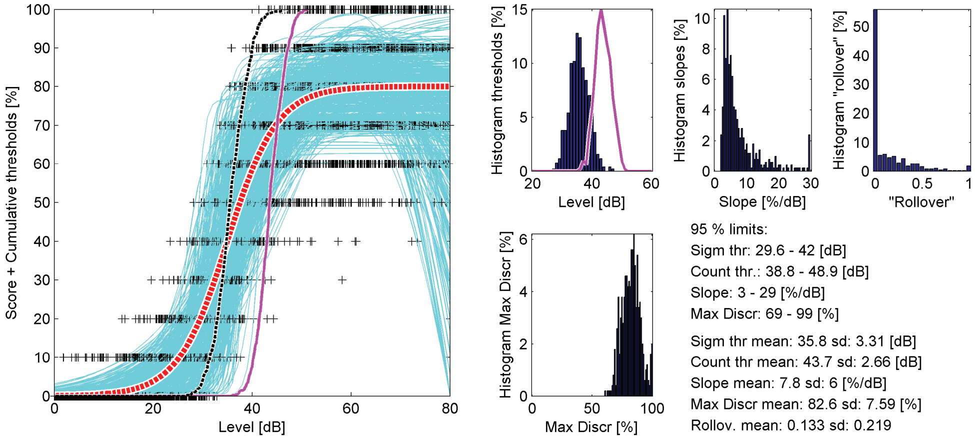


Figure 5.41 Results for HS3. Refer to Figure 5.39 for further explanation.

measuring between 5-75 dB for ten-word sets the interval between the levels had to be fixed at 7 dB. 110 items were used for each measurement. For the curve-fitting routine the standard deviation of the thresholds increases from 1.4 dB for HS1 to 2.9-3.3 dB for HS2-4. The counting routine results in lower standard deviations for the thresholds, but the mean values for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are not very accurate and the variability is large.

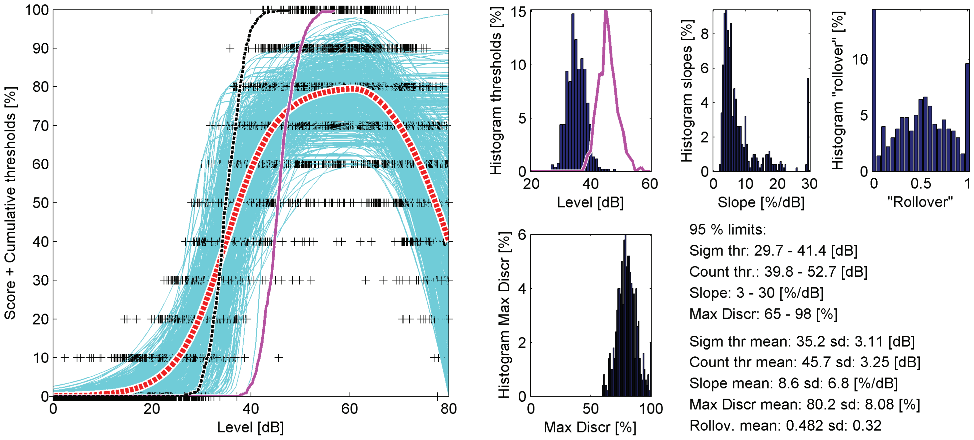


Figure 5.42 Results for HS4. Refer to Figure 5.39 for further explanation.

5.4.6.4 30-word sets at 17 dB intervals, 150 words/session

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in Figures 5.43-5.46. In order to achieve results within the ±7.5 dB range for 30-word sets the interval between the levels had to be fixed at 17 dB, but the simulations were performed between 1-69 dB in order to ensure that one of the levels was close to the threshold. 150 items were used for each measurement. For the curve-fitting routine the standard deviation of the thresholds increases from 1.2 dB for HS1 to 2.5-3.0

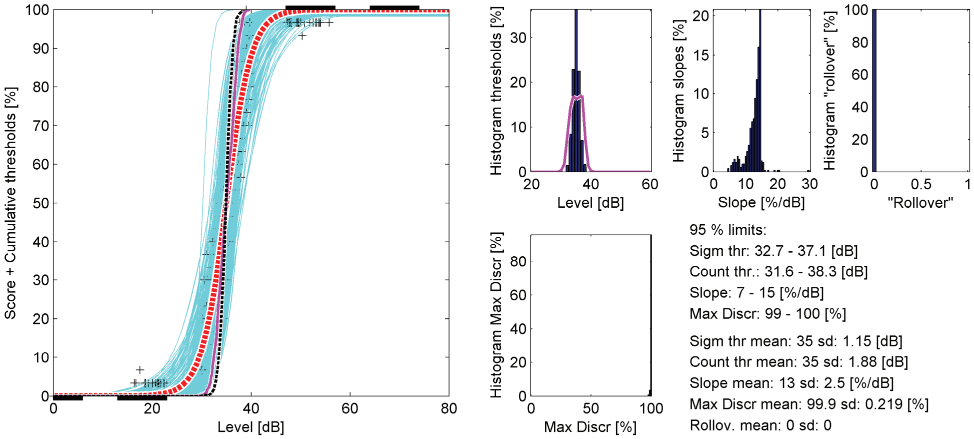


Figure 5.43 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 30 test items in each set measured at every level from 1-69 dB in 17 dB intervals. Refer to Figure 5.31 for explanation of the remaining details.

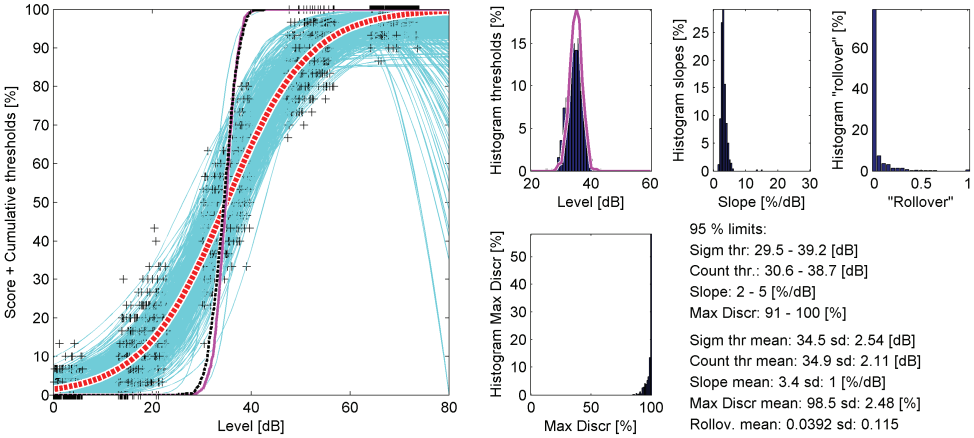


Figure 5.44 Results for HS2. Refer to Figure 5.43 for further explanation.

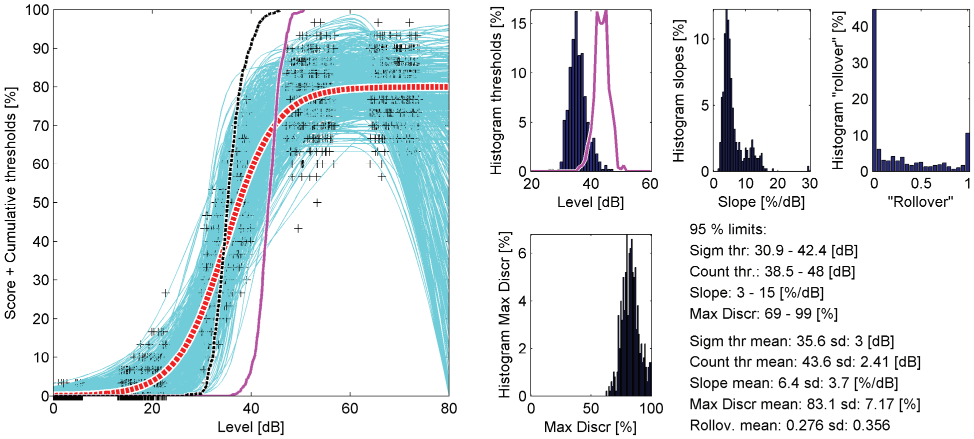


Figure 5.45 Results for HS3. Refer to Figure 5.43 for further explanation.

dB for HS2-4. The counting routine gave inaccurate results except for HS2. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are more accurate and the variability is smaller than the measurements in the preceding sections which involved fewer words at each level.

Simulations were also performed for the 16 %/dB slope as expected with three-word utterances in noise and with the rest of the parameters kept as before for HS1. These simulations produced the following results: Sigmoid thresholds: 95% 31.2-37.2 dB, mean 34.8 dB and standard deviation 1.29 dB. Count thresholds: 95% 30.3-37.2 dB, mean 35 dB and standard deviation 2.95 dB. These results could be much improved if we used the same number of test levels with closer intervals.

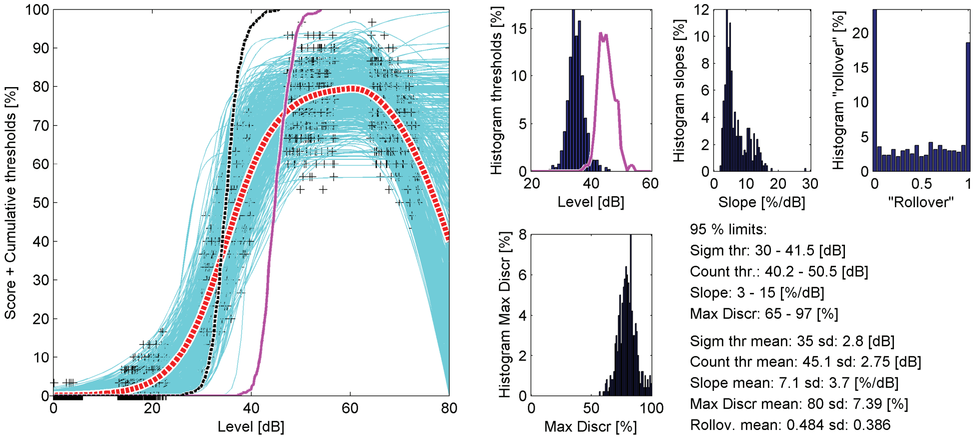


Figure 5.46 Results for HS4. Refer to Figure 5.43 for further explanation.

5.4.6.5 50-word sets at 21 dB intervals, 200 words/session

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in Figures 5.47-5.50. In order to achieve results within the ±7.5 dB range when using fifty-word sets the interval between the levels had to be 21 dB, but the simulations were performed in the 14-77 dB range in order to ensure that

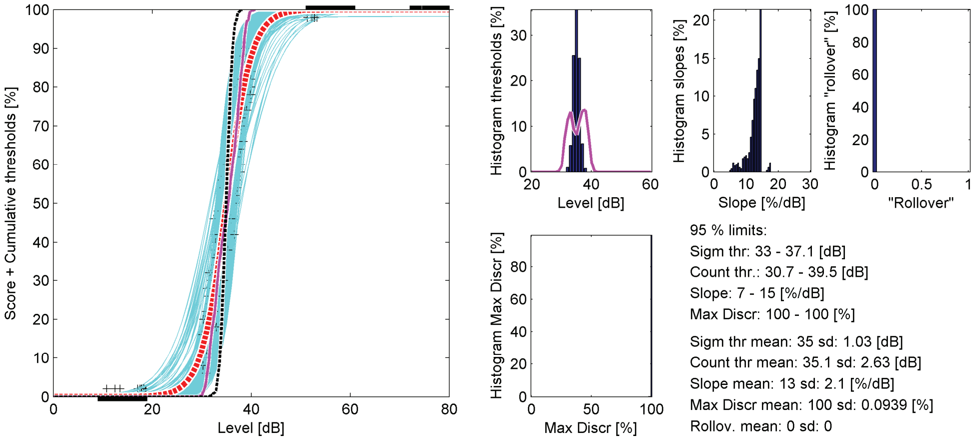


Figure 5.47 Simulated measurements of hypothetical subject HS1 according to the constant stimuli method. Results are calculated by both the curve-fitting routine and the counting routine. 50 test items in each set measured at every level from 14 - 78 dB in 21 dB intervals. Refer to Figure 5.31 for explanation of the remaining details.

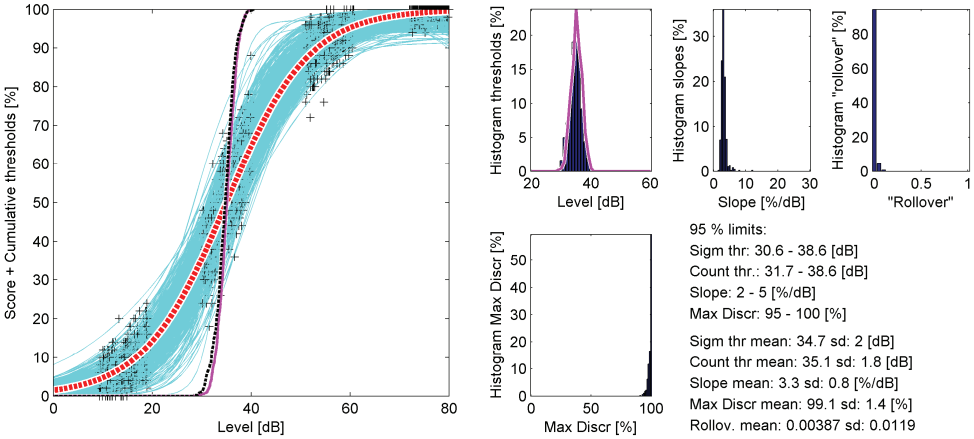


Figure 5.48 Results for HS2. Refer to Figure 5.47 for further explanation.

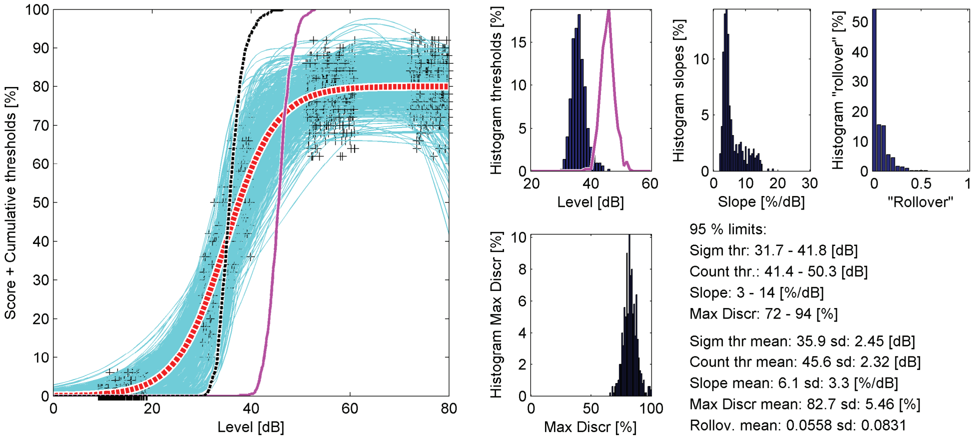


Figure 5.49 Results for HS3. Refer to Figure 5.47 for further explanation.

one of the levels would be close to the threshold. 200 items were used for each measurement. For the curve-fitting routine the standard deviation of the thresholds increases from 1.0 dB for HS1 to 2.0-2.5 dB for HS2-4. The counting routine gave inaccurate results except for HS2. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, and the standard deviation values are a little better for HS3-HS4 than in the preceding sections. The estimated slopes are also more accurate here, and the variability smaller, in line with the findings for the 30-words sets described in the preceding section.

Simulations were also performed for the 14 %/dB slope as expected with five-word sentences in noise and with the rest of the parameters kept as before for HS1. These simulations produced the following results: Sigmoid thresholds: 95% 32.7-36.4 dB, mean 34.9 dB and standard deviation 0.994 dB. Count thresholds: 95% 29.6-40.4 dB, mean 35.2 dB and standard deviation 3.55 dB.

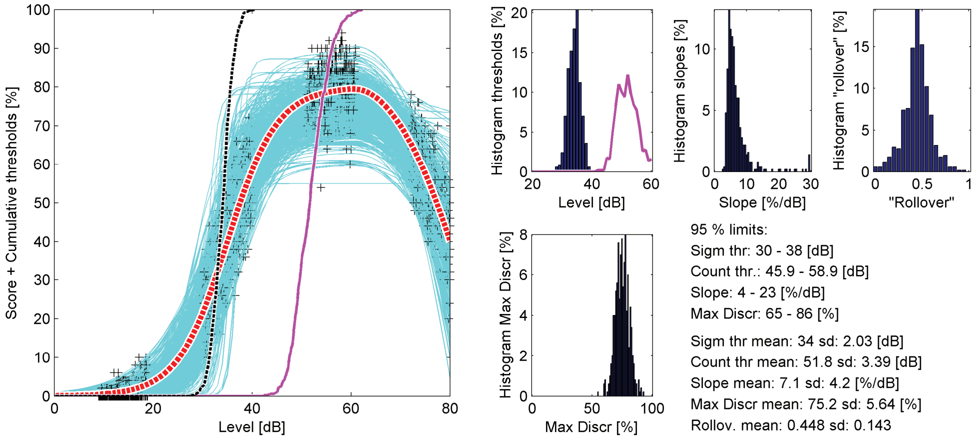


Figure 5.50 Results for HS4. Refer to Figure 5.47 for further explanation.

5.4.7 Summary of simulations presented in “HiST taleaudiometri”

The simulations made for the recommended measurement methods in “HiST taleaudiometri” are described in the report (Øygarden 2009) accompanying the CDs and DVD. In this section a summary of the results is presented.

5.4.7.1 The constant stimuli method with 10 words at 5 dB intervals (150 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 25-28 in “HiST taleaudiometri”.

The simulations demonstrate what to expect when we try to measure the performance-intensity curve with 10 words at each level in 5 dB steps. The chosen levels are between 10 and 80 dB.

For the curve-fitting routine the standard deviation of the thresholds increases from 1.2 dB for HS1 to 2.5-3.0 dB for HS2-HS4. The counting routine results in lower standard deviations for the thresholds except for HS4, but the mean values produced for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are not very accurate and the variability is large.

Two extra simulation sets were performed for a hypothetical subject with an expected normal slope for monosyllabic words of 7 %/dB. The first set, where the levels were adjusted in 5 dB intervals, showed a 1.5 dB standard deviation of thresholds, and 95 % of the thresholds were within ±3.0 dB of the correct result. The second set, where the levels were adjusted in 10 dB intervals, showed a 2.0 dB standard deviation of thresholds, and 95 % of the thresholds were within ±4.4 dB

5.4.7.2 The 80-20 % adaptive stimuli method with 30 words at 5 dB intervals

The results for the adaptive stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 33-36 in “HiST taleaudiometri”.

These simulations are the result of using the adaptive stimuli method with the lists of three-word utterances. For the purpose of reducing measurement time it is possible to confine the measuring to the range between 80 % and 20 % scores in 5 dB intervals. The threshold is estimated by drawing the performance intensity curve.

For the curve-fitting routine the standard deviation of the thresholds increases from 0.74 dB for HS1 to 2.2-2.4 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are located between +5.5 and -5.6 dB of the correct threshold. The counting routine produces lower standard deviations for the thresholds except for HS4, but the resulting mean values for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameters are good for HS1, but for HS2-HS4 the standard deviations are rather wide. The estimated slopes are close to correct and the standard deviations are lower than 2.9 %/dB.

The mean required number of items tested increased from 154 for HS1 to about 200 for HS2-HS4, which means that using 5-7 test lists is usually sufficient.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 16 %/dB for three-word utterances in noise. The standard deviation of the threshold was reduced to 0.55 dB for the curve-fitting routine.

5.4.7.3 The constant stimuli method with 3 words at 1.5 dB intervals (90 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 41-43 in "HiST taleaudiometri".

These simulations are the result of using test lists that are called quick-speed tests in "HiST taleaudiometri". The test lists consist of 30 three-word utterances where the level is reduced by 1.5 dB for each utterance. The threshold is estimated by the counting method.

For the counting routine the standard deviation of the thresholds increases from 1.1 dB for HS1 to 1.9-2.1 dB for HS2-4. 95 % of the estimated thresholds are found within +7.5 to -3.4 dB of the correct threshold for HS1-HS4.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 16 %/dB for three-word utterances in noise. The standard deviation of the threshold was reduced to 0.84 dB for the counting routine.

5.4.7.4 The 80-20 % adaptive stimuli method with 50 words at 10 dB intervals

The results for the adaptive stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 47-50 in "HiST taleaudiometri".

These simulations are the result of using the adaptive stimuli method with the lists of five-word sentences. For the purpose of reducing measurement time it is possible to confine the measuring to the range between 80 % and 20 % scores in 10 dB intervals. The threshold is estimated by drawing the performance intensity curve.

For the curve-fitting routine the standard deviation of the thresholds increases from 0.83 dB for HS1 to 2.1-2.3 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are found within +5.2 to -5.0 dB of the correct threshold. The counting routine results in lower standard deviations for the thresholds except for HS3-HS4, where the mean values produced are also wrong. The estimated maximum discrimination and rollover parameters are good for HS1, but for HS2-HS4 the standard deviations are rather wide. The estimated slopes are close to correct and the standard deviations are lower than 2.4 %/dB.

The mean required number of items tested increased from 158 for HS1 to about 210-235 for HS2-HS4, which means that using 3-5 test lists is usually sufficient.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 14 %/dB for five-word sentences in noise. The standard deviation of the threshold was reduced to 0.81 dB for the curve-fitting routine.

5.4.7.5 The constant stimuli method with 5 words at 2.5 dB intervals (100 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 55-57 in “HiST taleaudiometri”.

These simulations are the result of using test lists that are called quick-speed tests in “HiST taleaudiometri”. The test lists consist of 20 five-word sentences where the level is reduced by 2.5 dB for each sentence. The threshold is estimated by the counting method.

For the counting routine the standard deviation of the thresholds increases from 1.1 dB for HS1 to 2.0-2.1 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are located within +6.8 to -3.6 dB of the correct threshold.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 14 %/dB for five-word sentences in noise. The standard deviation of the threshold was reduced to 0.93 dB for the counting routine.

5.4.7.6 The constant stimuli method with 5 words at 2.5 dB intervals (50 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 64-66 in “HiST taleaudiometri”.

These simulations are the result of using test lists that are similar to the quick-speed tests in “HiST taleaudiometri”. The test lists consist of 10 five-word sentences where the level is reduced by 2.5 dB for each sentence. The threshold is estimated by the counting method. These types of list were developed for tests with masking noise, such as the binaural test for earphones and the free-field audio DVD test with surround sound.

For the counting routine the standard deviation of the thresholds increases from 1.1 dB for HS1 to 1.8-2.1 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are found within +3.9 to -3.9 dB of the correct threshold.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 14 %/dB for five-word sentences in noise. The standard deviation of the threshold was reduced to 0.93 dB for the counting routine.

5.4.7.7 The 100-0 % adaptive stimuli method with 9 words at 5 dB intervals

The results for the adaptive stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 68-71 in "HiST taleaudiometri".

These simulations are the result of using the adaptive stimuli method with the monosyllabic numerals (digit triplets). Nine numerals (three digit triplets) are presented at each level. Scores should be obtained for all the levels in 5 dB intervals between a level giving the top score of 100 % and a level with 0 % score. The threshold is estimated by drawing the performance intensity curve.

For the curve-fitting routine the standard deviation of the thresholds increases from 1.4 dB for HS1 to 3.0-3.4 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are found within +7.9 to -6.8 dB of the correct threshold. The counting routine produces lower standard deviations for the thresholds except for HS4, but the mean values for HS3-HS4 are wrong. The estimated maximum discrimination and rollover parameters are good for HS1-HS2, but the standard deviations for HS3-HS4 are rather wide. The estimated slopes are not very accurate and the variability is large.

The mean required number of items tested increased from 42 for HS1 to about 80-94 for HS2-HS4, which means that using 14-32 triplets is usually sufficient.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 17 %/dB for monosyllabic numerals. The standard deviation of the threshold was reduced to 0.99 dB for the curve-fitting routine.

5.4.7.8 The constant stimuli method with 3 words at 2 dB intervals (90 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 81-83 in "HiST taleaudiometri".

These simulations are the result of using test lists that are similar to the quick-speed tests in "HiST taleaudiometri". The test lists consist of 30 three-word utterances where the level is reduced by 2 dB for each utterance. The threshold is estimated by the counting method. These lists were developed for the test without masking noise for measuring the free-field speech recognition threshold on the audio DVD.

For the counting routine the standard deviation of the thresholds increases from 1.3 dB for HS1 to 1.9-2.4 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are located within +8.0 to -3.8 dB of the correct threshold.

5.4.7.9 The constant stimuli method with 5 words at 3 dB intervals (100 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 83-85 in “HiST taleaudiometri”.

These simulations are the result of using test lists that are similar to the quick-speed tests in “HiST taleaudiometri”. The test lists consist of 20 five-word sentences where the level is reduced by 3 dB for each sentence. The threshold is estimated by the counting method. These lists were developed for the test without masking noise for measuring the free-field speech recognition threshold on the audio DVD.

For the counting routine the standard deviation of the thresholds increases from 1.2 dB for HS1 to 2.1-2.3 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are within +8.2 to -4.6 dB of the correct threshold.

5.4.7.10 The constant stimuli method with 5 words at 0.75 dB intervals (150 words/session)

The results for the constant stimuli method with estimates calculated by both the curve-fitting routine and the counting routine are displayed in figures on pages 99-101 in “HiST taleaudiometri”.

These simulations are the result of using test lists that are similar to the quick-speed tests in “HiST taleaudiometri”. The test lists consist of 30 five-word sentences where the level is reduced by 0.75 dB for each sentence. The threshold is estimated by the counting method. These lists were developed for the test with four channels of uncorrelated masking noise for measuring the free-field signal-to-noise ratio speech recognition threshold on the audio DVD with surround sound.

For the counting routine the standard deviation of the thresholds increases from 0.59 dB for HS1 to 1.3-1.5 dB for HS2-4. For HS1-HS4 95 % of the estimated thresholds are located within +2.8 to -3.0 dB of the correct threshold.

An extra simulation was performed for a hypothetical subject with an expected normal slope of 14 %/dB for five-word sentences in noise. The standard deviation of the threshold was reduced to 0.54 dB for the counting routine.

5.4.8 Detection of differences in speech recognition thresholds

5.4.8.1 Introduction

One important purpose of performing speech audiometry is to detect differences in the speech recognition thresholds between different treatments. We may be interested in deciding which results are best when comparing two different hearing aids, two different adjustments on hearing aids or cochlear implants, two different subjects or speech presented in noise with different characteristics, etc.

To illustrate the problem connected with measuring differences with optimal accuracy the results of some simulations are presented in Figures 5.51-5.53. All simulations are based on measurements according to the constant stimuli method with 10-word sets at 5 dB intervals between 20 and 80 dB. The thresholds estimated with the curve-fitting routine had a standard deviation of sd_{thr} = 1.25 dB for hypothetical subject HS1, with a slope of 10 %/dB, 100 % maximum speech recognition score and no rollover. 500 simulations were performed for each measurement condition. The statistics are consequently based on the 250 000 differences that can be calculated between two sets of 500 thresholds.

For independent variables we know that the standard deviation of differences or sums between variables with the same standard deviations is √2 larger than the standard deviation for each part. We would thus expect the standard deviation for the difference in sigmoid thresholds in Figure 5.51 to become $\sqrt{2}\cdot sd_{thr} = \sqrt{2}\cdot 1.25 = 1.77$ which corresponds well with sd_{diff} = 1.79 in the figure (Sigm … sd: 1.79 [dB]). The 5 % upper limit (one-tailed) of this distribution of differences is $sd_{\mathrm{diff}}\cdot 1.65 = 1.79\cdot 1.65 = 2.95$ dB when the distribution is normal. This also corresponds well with the upper 90 % limit (two-tailed) of the sigmoid threshold of 2.9 dB in Figure 5.51. The conclusion we can draw from this figure is that if we measure differences in threshold between two test situations larger than $\sqrt{2}\cdot sd_{thr}\cdot 1.65$, this is an indication that the two test situations are not equal since we would only get such results in 5 % of the instances given equal test situations.

In Figure 5.52 the difference in the thresholds for the hypothetical subjects is $\sqrt{2}\cdot sd_{thr}\cdot 1.65 = 2.95$. From the cumulative distribution in the large panel (dashed line) we see that 5 % of the differences are below 0 dB. That is with such a difference between the hypothetical subjects the measured differences would be correctly sorted in 95 % of the instances. But in order to be certain that we detected real differences in 95 % of the instances we need even greater difference between the two hypothetical subjects as described in the text for Figure 5.53.

In Figure 5.53 the difference between the two hypothetical subjects threshold is $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.65 = 4.4\cdot sd_{thr}=$ 5.9 dB. This means that approximately 95 % of the simulated differences are greater than 95 % of the differences in Figure 5.51. The lower 90 % limit of the sigmoid thresholds (2.8 dB) in Figure 5.53 should consequently be the same as the upper 90 % limit of the sigmoid thresholds (2.9 dB) in Figure 5.51. However we have some inexactness from the simulations.

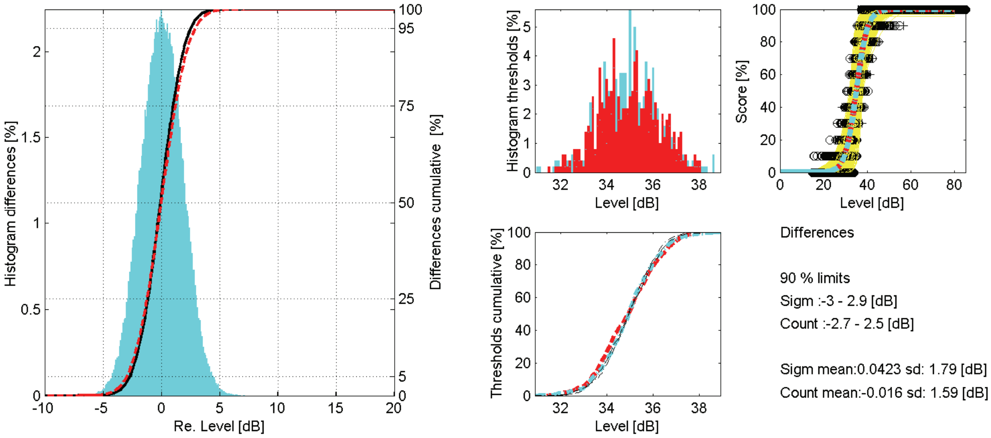


Figure 5.51 Differences between two simulated measurements on two hypothetical subjects with identical threshold of 35.0 dB. The upper right panel shows the performance-intensity curves for the hypothetical subjects with simulated responses and fitted curves for one of the subjects. The middle top panel shows overlapping histograms of the thresholds for the two subjects. The lower panel shows the cumulative thresholds for the two subjects. The large panel shows a histogram of the differences between the estimated thresholds for the two subjects and the cumulative distribution of the differences estimated by the curve-fitting routine, dashed line; and the counting routine, solid line.

The preceding judgements have all used 95 % security for a one-tailed test which requires 1.65·sd difference. The use of one-tailed tests is only appropriate when we expect that one particular treatment is better than the other. When comparing different hearing aids, we do not know which aid is the better and we have to use a two-tailed test. The required difference for 95 % security will then be 1.96·sd. Then the required difference between measurements is $\sqrt{2}\cdot sd_{thr}\cdot 1.96$ to get an indication that the two test situations are not equal since we would only get such results in 5 % of the instances given equal test situations. The true differences between two treatments need to be $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.96$ to be able to register the required difference of $\sqrt{2}\cdot sd_{thr}\cdot 1.96$ in 95 % of the cases.

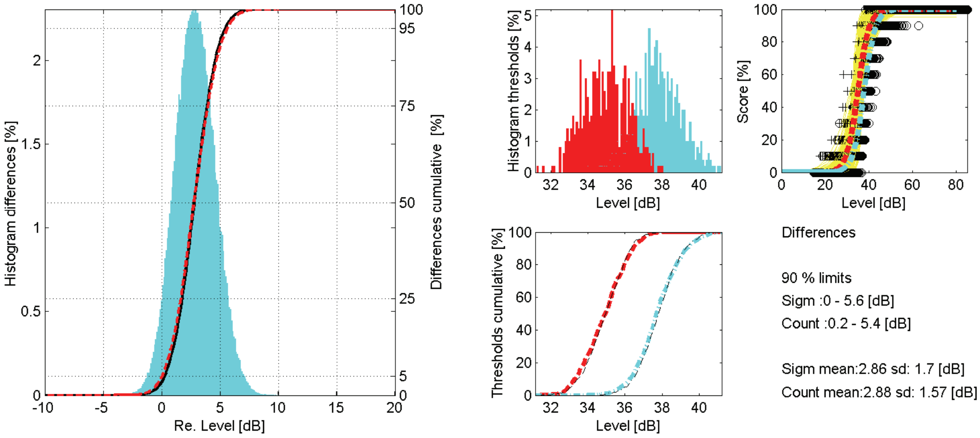


Figure 5.52 Differences between two simulated measurements on two hypothetical subjects with thresholds of 35.00 and 37.95 dB. Refer to Figure 5.51 for further explanation.

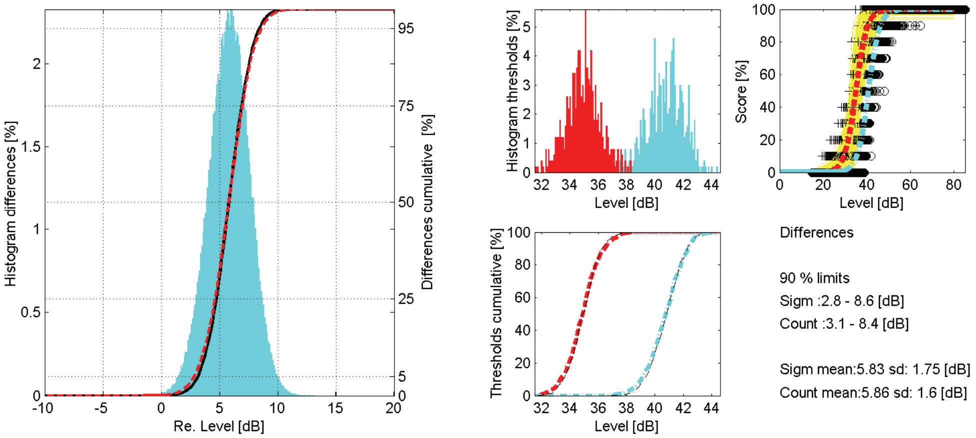


Figure 5.53 Differences between two simulated measurements on two hypothetical subjects with thresholds of 35.0 and 40.9 dB. Refer to Figure 5.51 for further explanation.

Since $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.65$ (one-tailed) or $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.96$ (two-tailed) is the true difference between two treatments needed to be registered in 95 % of the cases, we need speech audiometry tests with low standard deviations of the thresholds in order to accurately measure that one treatment is different from another. In section 5.4.8.2 standard deviations from the simulation results in section 5.4 will be presented, as well as the limits$\sqrt{2}\cdot 1.65\cdot sd_{thr}$, $2\cdot\sqrt{2}\cdot 1.65\cdot sd_{thr}$ and $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.96$.

The $\sqrt{2}\cdot 1.65\cdot sd_{thr}65$ (one-tailed) or $\sqrt{2}\cdot sd_{thr}\cdot 1.96$ (two-tailed) criterion can be used for two evaluations. First, if we measure differences larger than this criterion, then we know that this would only occur in 5 % of the cases given that the thresholds for the treatments are truly equal. Second, if the difference between the true thresholds for the treatments is $\sqrt{2}\cdot 1.65\cdot sd_{thr}$, the measured differences would be sorted correctly in 95 % of the cases.

The $2\cdot\sqrt{2}\cdot 1.65\cdot sd_{thr}$ (one-tailed) or $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.96$ (two-tailed) criterion is the required difference between the true thresholds of two treatments allowing us to measure different thresholds in 95 % of the cases.

The value of the standard deviations of the thresholds from the simulations will have large variations and will be influenced by many factors. Among these factors is, first, the slope of the performance-intensity function. A steep slope gives a lower standard deviation of the threshold for most of the measurement methods. The steepness of slopes was discussed more comprehensively in section 5.1.1. Second, the standard deviation of the thresholds will also be influenced by the listener's maximum speech recognition score in relation to the specific speech material used. Different measurement methods will be more or less affected by reduced maximum speech recognition score. Finally, rollover will also vary both for the different listeners and the different speech material combined. The measurement methods will be more or less insensitive to this effect.

5.4.8.2 Detection of differences in SRTs for the simulated speech audiometry procedures

The results of the simulations in sections 5.4.1-5.4.7 are here used as described in the preceding section to estimate the smallest difference between treatments needed in order to determine whether they represent identical or different treatments. The results are presented in Table 5.5. Table 5.5 shows that measuring subtle differences between treatments involving speech audiometry can be difficult unless we select the test material and the testing procedure carefully. We are of course unable to control the hearing performance of our test subject. If the performance-intensity curve is similar to that of our hypothetical subjects HS2-HS4, only a few of the procedures will manage to discern differences between treatments lower than 10 dB (in the $2\cdot\sqrt{2}\cdot 1.65\cdot sd$, one-tailed or $2\cdot\sqrt{2}\cdot sd_{thr}\cdot 1.96$, two-tailed columns): only the Brand and Kollmeier A2 threshold and slope method and the counting methods used with five-word sentences at 2.5 or 0.75 dB intervals or with the three-word utterances at 1.5 dB intervals satisfy this requirement. A 10 dB difference between treatments is a large one, and when measuring differences between hearing aids or performance differences in various noise situations we need to be able to discern smaller differences.

Table 5.5 Computed required differences in the thresholds based on the standard deviation from the simulations in section 5.4.1-5.4.7. The $\sqrt{2}\cdot 1.65\cdot$sd column is the required difference between two situations for sorting them correctly. The $2\cdot\sqrt{2}\cdot 1.65\cdot$sd (one-tailed) and the $2\cdot\sqrt{2}\cdot 1.96\cdot$sd (two-tailed) column is the required difference between two situations for registering that they are different with 95 % confidence .

Fig/Sec	HS	SD	$\sqrt{2}\cdot$ 1.65·sd	$2\cdot\sqrt{2}\cdot$ 1.65·sd	$2\cdot\sqrt{2}\cdot$ 1.96·sd
ISO 8253-3 Threshold level procedure A					
5.4.1.1	**10 items in test set**				
5.7	HS1	1.4	3.3	6.5	7.8
5.8	HS2	4.0	9.3	18.6	22.1
5.9	HS3	5.4	12.6	25.2	30.0
5.10	HS4	5.5	12.8	25.6	30.4
5.4.1.2	**23 items in test set**				
5.11	HS1	1.0	2.3	4.5	5.4
5.12	HS2	2.7	6.2	12.4	14.7
5.13	HS3	2.4	5.7	11.4	13.5
5.14	HS4	2.5	5.8	11.6	13.7
text 30w	16%/dB	0.7	1.6	3.1	3.7
text 50w	14%/dB	0.6	1.3	2.7	3.2
5.4.2	**The Hagerman and Kinnefors S/N-threshold method**				
5.15	HS1	0.7	1.6	3.2	3.8
5.16	HS2	1.9	4.5	8.9	10.6
5.17	HS3	1.9	4.4	8.7	10.4
5.18	HS4	2.0	4.6	9.1	10.9
text	14%/dB	0.5	1.2	2.4	2.8
5.4.3	**The Hagerman and Kinnefors SRT-threshold method**				
5.19	HS1	1.0	2.3	4.7	5.5
5.20	HS2	2.7	6.4	12.8	15.2
5.21	HS3	3.6	8.4	16.9	20.1
5.22	HS4	3.7	8.7	17.4	20.6
text	14%/dB	0.8	1.8	3.6	4.3
5.4.4	**The Brand and Kollmeier A1 threshold method**				
5.23	HS1	0.6	1.3	2.6	3.1
5.24	HS2	2.7	6.2	12.4	14.7
5.25	HS3	2.4	5.6	11.3	13.4
5.26	HS4	2.2	5.2	10.4	12.4
text	14%/dB	0.4	0.9	1.9	2.3
5.4.5	**The Brand and Kollmeier A2 threshold and slope method**				
5.27	HS1	0.7	1.6	3.1	3.7
5.28	HS2	1.8	4.1	8.2	9.8
5.29	HS3	1.9	4.4	8.7	10.4
5.30	HS4	2.1	4.8	9.6	11.4
text	14%/dB	0.5	1.1	2.3	2.7

Table 5.5 continued

The performance-intensity function procedure and the counting procedure

			√2·	2·√2·	2·√2·	Count	√2·	2·√2·	2·√2·
Fig/Sec	**HS**	**sd**	**1.65·sd**	**1.65·sd**	**1.96·sd**	**sd**	**1.65·sd**	**1.65·sd**	**1.96·sd**
5.4.6.1	**3 word sets at 1.5 dB intervals, 141 words/session**								
5.31	HS1	1.2	2.9	5.7	6.8	1.1	2.5	5.1	6.0
5.32	HS2	2.7	6.2	12.4	14.7	2.1	4.9	9.9	11.8
5.33	HS3	2.8	6.6	13.2	15.6	2.1	4.8	9.7	11.5
5.34	HS4	2.9	6.7	13.3	15.9	2.1	4.8	9.7	11.5
text	16%/dB	0.9	2.2	4.4	5.2	0.8	1.9	3.9	4.6
5.4.6.2	**5 word sets at 2.5 dB intervals, 145 words/session**								
5.35	HS1	1.2	2.8	5.6	6.6	1.1	2.5	5.0	6.0
5.36	HS2	2.5	5.9	11.9	14.1	2.0	4.7	9.5	11.3
5.37	HS3	2.8	6.6	13.2	15.6	2.3	5.3	10.6	12.6
5.38	HS4	2.9	6.7	13.4	15.9	2.6	6.1	12.2	14.5
text	14%/dB	1.1	2.6	5.1	6.1	1.0	2.3	4.6	5.5
5.4.6.3	**10 word sets at 7 dB intervals, 110 words/session**								
5.39	HS1	1.4	3.2	6.3	7.5	1.3	3.0	6.1	7.2
5.40	HS2	2.9	6.7	13.4	15.9	2.3	5.5	10.9	13.0
5.41	HS3	3.3	7.7	15.4	18.3	2.7	6.2	12.4	14.7
5.42	HS4	3.1	7.3	14.5	17.2	3.3	7.6	15.2	18.0
5.4.6.4	**30 word sets at 17 dB intervals, 150 words/session**								
5.43	HS1	1.2	2.7	5.4	6.4	1.9	4.4	8.8	10.4
5.44	HS2	2.5	5.9	11.9	14.1	2.1	4.9	9.8	11.7
5.45	HS3	3.0	7.0	14.0	16.6	2.4	5.6	11.2	13.4
5.46	HS4	2.8	6.5	13.1	15.5	2.8	6.4	12.8	15.2
text	16%/dB	1.3	3.0	6.0	7.2	3.3	7.6	15.2	18.1
5.4.6.5	**50 word sets at 21 dB intervals, 200 words/session**								
5.47	HS1	1.0	2.4	4.8	5.7	2.6	6.1	12.3	14.6
5.48	HS2	2.0	4.7	9.3	11.1	1.8	4.2	8.4	10.0
5.49	HS3	2.5	5.7	11.4	13.6	2.3	5.4	10.8	12.9
5.50	HS4	2.0	4.7	9.5	11.3	3.4	7.9	15.8	18.8
text	14%/dB	1.0	2.3	4.6	5.5	3.6	8.3	16.6	19.7
Summary of simulations presented in "HiST taleaudiometri"									
5.4.7.1	**10 word sets at 5 dB intervals, 150 words/session**								
3	HS1	1.2	2.8	5.6	6.6	1.1	2.5	4.9	5.9
4	HS2	2.5	5.8	11.6	13.7	2.1	4.9	9.8	11.6
5	HS3	3.0	6.9	13.8	16.4	2.4	5.6	11.2	13.4
6	HS4	2.7	6.3	12.6	15.0	3.2	7.5	15.0	17.8
7	7%/dB 5	1.5	3.4	6.8	8.0	1.3	3.1	6.2	7.3
8	7%/dB10	2.0	4.8	9.5	11.3	1.9	4.4	8.8	10.4
5.4.7.2	**30 word sets at 5 dB intervals 80-20% method**								
9	HS1	0.7	1.7	3.5	4.1	0.7	1.6	3.2	3.8
10	HS2	2.2	5.0	10.1	12.0	1.3	3.1	6.1	7.3
11	HS3	2.4	5.5	11.0	13.0	2.0	4.7	9.4	11.1
12	HS4	2.4	5.5	11.0	13.0	2.9	6.7	13.5	16.0
13	16%/dB	0.6	1.3	2.6	3.1	0.5	1.2	2.4	2.9

Table 5.5 continued.

The performance-intensity function procedure and the counting procedure									
			√2·	2·√2·	2·√2·	Count	√2·	2·√2·	2·√2·
Fig/Sec	HS	sd	1.65·sd	1.65·sd	1.96·sd	sd	1.65·sd	1.65·sd	1.96·sd
5.4.7.3	**3 word sets at 1.5 dB intervals 90 words/session**								
14	HS1	1.3	2.9	5.9	7.0	1.1	2.6	5.2	6.2
15	HS2	2.9	6.7	13.3	15.8	1.9	4.4	8.9	10.5
16	HS3	3.2	7.5	15.1	17.9	2.0	4.6	9.2	10.9
17	HS4	3.0	7.0	14.1	16.7	2.1	4.9	9.9	11.8
18	16%/dB	0.9	2.2	4.4	5.2	0.8	2.0	3.9	4.6
5.4.7.4	**50 word sets at 10 dB intervals 80-20% method**								
19	HS1	0.8	1.9	3.9	4.6	0.7	1.6	3.3	3.9
20	HS2	2.0	4.8	9.5	11.3	1.3	3.0	6.0	7.2
21	HS3	2.1	5.0	10.0	11.9	2.7	6.3	12.6	15.0
22	HS4	2.3	5.3	10.5	12.5	4.2	9.7	19.5	23.1
23	14%/dB	0.8	1.9	3.8	4.5	0.7	1.6	3.3	3.9
5.4.7.5	**5 word sets at 2.5 dB intervals, 100 words/session**								
24	HS1	1.3	3.0	6.0	7.1	1.1	2.6	5.3	6.3
25	HS2	2.7	6.3	12.6	15.0	2.0	4.6	9.2	11.0
26	HS3	3.2	7.5	15.0	17.8	2.1	4.8	9.6	11.4
27	HS4	3.3	7.6	15.2	18.1	2.0	4.8	9.5	11.3
28	14%/dB	1.1	2.5	4.9	5.8	0.9	2.2	4.3	5.2
5.4.7.6	**5 word sets at 2.5 dB intervals, 50 words/session**								
29	HS1	1.4	3.3	6.7	7.9	1.1	2.5	5.0	6.0
30	HS2	4.2	9.8	19.6	23.2	2.1	4.8	9.6	11.4
31	HS3	3.8	8.9	17.7	21.1	1.8	4.1	8.2	9.7
32	HS4	3.9	9.0	18.0	21.4	1.8	4.3	8.5	10.1
33	14%/dB	1.1	2.5	5.1	6.0	0.9	2.2	4.3	5.1
5.4.7.7	**9 word sets at 5 dB intervals 100-0% method**								
34	HS1	1.4	3.2	6.3	7.5	1.2	2.8	5.6	6.7
35	HS2	3.0	6.9	13.8	16.4	2.4	5.6	11.2	13.3
36	HS3	3.2	7.4	14.9	17.7	2.8	6.5	13.0	15.5
37	HS4	3.4	7.9	15.8	18.7	3.4	7.9	15.8	18.7
38	17%/dB	1.0	2.3	4.6	5.5	0.9	2.2	4.4	5.2
5.4.7.8	**3 word sets at 2 dB intervals 90 words/session**								
40	HS1	1.5	3.5	7.0	8.3	1.3	3.0	6.0	7.2
41	HS2	2.3	5.4	10.8	12.8	1.9	4.5	9.0	10.7
42	HS3	3.5	8.3	16.5	19.6	2.4	5.5	11.1	13.1
43	HS4	3.5	8.2	16.5	19.6	2.4	5.6	11.2	13.3
5.4.7.9	**5 word sets at 3 dB intervals 100 words/session**								
44	HS1	1.5	3.4	6.8	8.0	1.2	2.8	5.6	6.7
45	HS2	3.1	7.2	14.4	17.1	2.1	5.0	9.9	11.8
46	HS3	3.3	7.7	15.4	18.3	2.3	5.4	10.8	12.8
47	HS4	3.5	8.1	16.3	19.3	2.3	5.3	10.7	12.7
5.4.7.10	**5 word sets at 0.75 intervals, 150 words/session**								
48	HS1	0.8	1.8	3.7	4.3	0.6	1.4	2.7	3.3
49	HS2	3.1	7.3	14.6	17.3	1.5	3.5	7.1	8.4
50	HS3	2.9	6.8	13.6	16.2	1.4	3.2	6.4	7.6
51	HS4	2.9	6.8	13.6	16.2	1.3	3.0	5.9	7.0
52	14%/dB	0.6	1.5	2.9	3.5	0.5	1.3	2.5	3.0

In section 5.1.1 the conclusions were that hypothetical subjects HS2-HS4 came close to representing worst-case scenarios for monosyllabic words. When we want to measure differences between hearing aid performances etc., the natural selection of material will be three-word utterances or five-word sentences, and the measurements must be performed with noise. The conclusion for this type of test was that both normal-hearing subjects and subjects with mild to moderate hearing loss had almost the same slope when measuring with stationary noise. The simulations with the 16 %/dB slopes for three-word utterances and 14 %/dB for five-word sentences need to be emphasized when judging the possibilities for finding small differences between treatments. Thus, the best procedures from our simulations would be those selected in Table 5.6.

Table 5.6 The best procedures for three-word utterances and five-word sentences.

Procedure	words	sd	√2·1.65·sd	2·√2·1.65·sd	2·√2·1.96·sd	mean items	sd items
ISO 8253-3 Threshold level procedure A	3	0.7	1.6	3.1	3.7	94	16
ISO 8253-3 Threshold level procedure A	5	0.6	1.3	2.7	3.2	167	30
The Hagerman and Kinnefors S/N-threshold method	5	0.5	1.2	2.4	2.8	100	0
The Hagerman and Kinnefors SRT-threshold method	5	0.8	1.8	3.6	4.3	60	4
The Brand and Kollmeier A1 threshold method	5	0.4	0.9	1.9	2.3	95	0
The Brand and Kollmeier A2 threshold and slope method	5	0.5	1.1	2.3	2.7	100	0
The 80-20 % adaptive stimuli method with the curve-fitting routine, 30-word sets at 5 dB intervals	3	0.6	1.3	2.6	3.1	147	11
The constant stimuli method with the counting routine, 3-word sets at 1.5 dB intervals 90 words/session	3	0.8	2.0	3.9	4.6	90	0
The 80-20 % adaptive stimuli method with the curve-fitting routine, 50-word sets at 10 dB intervals	5	0.8	1.9	3.8	4.5	153	11
The constant stimuli method with the counting routine, 5-word sets at 2.5 dB intervals, 50 words/session	5	0.9	2.2	4.3	5.1	50	0
The constant stimuli method with the counting routine, 5-word sets at 0.75 intervals, 150 words/session	5	0.5	1.3	2.5	3.0	150	0

It is difficult to compare the procedures in Table 5.6 because the number of items used for each measurement varies. New modified simulations were performed. The number of test items was normalized to around 150. To accomplish this some modification of intervals between levels or the number of sentences had to be made. The range between the highest and lowest test level was also changed for some of the procedures in order to reduce the standard deviation of the threshold. Furthermore, the chosen starting level was also changed for some procedures. Simulation with three-word utterances were performed with a slope of 16 %/dB and a slope of 14 %/dB was used for five-word sentences. The results of the modified procedures are presented in Table 5.7.

The range of standard deviations for the thresholds was now reduced to between 0.292-0.542 dB compared to the 0.4-0.9 dB values produced by the procedures in Table 5.6. Some of the differences between the procedures are commented on in the following paragraphs.

The ISO 8253-3 Threshold level procedure A variants were the worst performers both for three-word and the five-word material. One reason can be that even though many test sets are evaluated, only the last two scores are used to compute the threshold. The rest of the scores are used to adjust the levels during measurement, but are not considered when the threshold is estimated.

The Hagerman and Kinnefors methods performed very well. The S/N-threshold method was a little better than the SRT-method. These methods were developed for the slope in the Swedish Hagerman material for measurement in quiet and in noise respectively. The slopes for our five-word and three-word material differ from those in the Hagerman material, and a modified Hagerman and Kinnefors procedure for thee-word utterances was developed for these simulations. The 16 %/dB slope shows a very fine fit with the testing of 6 items, which means a pair of three-word utterances at each level (one item is 16.67 % of six items). If we set the target as a 50 % score then each word deviant from three recognized words out of the six in a set is used to adjust the level by one dB before the next test. This procedure had the lowest standard deviation of all the procedures evaluated.

The Brand and Kollmeier procedures were also very good, producing approximately the same standard deviations as the Hagerman and Kinnefors S/N and SRT-procedures.

The procedures with threshold estimations based on the curve-fitting routines and the counting routines lagged a little behind the best procedures but not as badly as the ISO 8253-3 procedures. The 80-20 % adaptive stimuli method used with the curve-fitting routine was best among these for 30-word sets, but for the 50-word sets it was the worst. The standard deviations of the two modified constant stimuli methods used with counting routines were found in between values for the two curve fitting routines.

Table 5.7 The best procedures for three-word utterances and five-word sentences modified for use with approximately 150 test items.

Procedures modified for use with about 150 test items	words	sd	√2·1.65·sd	2·√2·1.65·sd	2·√2·1.96·sd	mean items	sd items
ISO 8253-3 Threshold level procedure A Modification: Starting level 5 dB over threshold 30 words in set 2 dB intervals	3	0.51	1.2	2.4	2.8	150	50
ISO 8253-3 Threshold level procedure A Modification: Starting level 5 dB over threshold 50 words in 4dB intervals	5	0.542	1.3	2.5	3.0	153	30
The Hagerman and Kinnefors S/N-threshold method. Modification: 25 sentences in measurement	5	0.343	0.8	1.6	1.9	148	4
The Hagerman and Kinnefors SRT-threshold method. Modification: 25 sentences in measurement	5	0.382	0.9	1.8	2.1	147	4
New method: Modified Hagerman and Kinnefors threshold method for pairs of three-word utterances. Level change rule -3, -2, -1, 0, 1, 2 and 3 dB for scores 6, 5, 4, 3, 2, 1 and 0. 22 pairs in measurement	2·3	0.292	0.7	1.4	1.6	151	4
The Brand and Kollmeier A1 threshold method Modification: 30 sentences are evaluated	5	0.343	0.8	1.6	1.9	150	0
The Brand and Kollmeier A2 threshold and slope method Modification: 30 sentences are evaluated	5	0.394	0.9	1.8	2.2	150	0
The 80-20 % adaptive stimuli method with the curve-fitting routine, 30-word sets modified to starting level 5 dB over threshold at 2 dB intervals	3	0.401	0.9	1.9	2.3	151	27
The constant stimuli method with the counting routine, 3-word sets modified to measure 50 levels over 20 dB range. 150 words/session	3	0.464	1.1	2.2	2.6	150	0
The 80-20 % adaptive stimuli method with the curve-fitting routine, 50-word sets modified to starting level 5 dB over threshold at 5 dB intervals	5	0.502	1.2	2.3	2.8	151	15
The constant stimuli method with the counting routine, 5-word sets modified to measure 30 levels over 20 dB range. 150 words/session	5	0.464	1.1	2.2	2.6	150	
The constant stimuli method with the counting routine, 5 word sets at 0.75 intervals, 150 words/session. Not modified	5	0.537	1.2	2.3	2.3	150	0

Three-word utterance procedures performed a little better than five-word sentence procedures for ISO 8253-3, Hagerman and Kinnefors and curve-fitting routines, but not for the counting routine, where the standard deviations were the same for both procedures. The better performance of three-word procedures can probably be explained by the steeper slope of 16 %/dB, versus 14 %/dB for the five-word sentences.

Three of the procedures can be used for estimating the slopes. The best estimate was given by the Brand and Kollmeier A2 threshold procedure with a mean value of 15 %/dB and standard deviation of 2.4 %/dB. The 80-20 % adaptive stimuli method used with the curve-fitting routine for 50-word sets gave the same mean value, but the standard deviation increased to 3.5 %/dB. The correct slope should have been 14 %/dB for the five-word sentences. For the three-word utterances the correct slope should be 16 %/dB, but the 80-20 % adaptive stimuli method used with the curve-fitting routine for 30-word sets gave a mean value of 18 %/dB and a standard deviation of 4.3 %/dB. Thus the performance was a little worse than for the five-word procedures.

5.5 Discussions

5.5.1 Speech recognition threshold - SRT

In Chapter 3 the use of three-word utterances instead of spondee words for measuring the speech recognition threshold was discussed, and we decided to recommend using the three-word utterances for measuring SRT. The simulations in this chapter show that many existing procedures could be used for this type of measurements. The slope of the performance-intensity curve has been found to be 10 %/dB for normal-hearing subjects. The slope for hearing-impaired subjects with this material has not been established, but because each test list consists of the same well-known 30 words we do not expect the slope to deteriorate much for subjects with mild to moderate sensorineural impairment. Some of the procedures selected for detecting differences in section 5.4.8.2 may also be feasible for measuring SRT, but in order to get an impression of how well they function we need to evaluate the standard deviations of the thresholds for hypothetical subject HS1, who had a slope of 10 %/dB. The following procedures need to be evaluated:

1. The ISO 8253-3 threshold level procedure A can be excluded because it performed worse than the other procedures.
2. The modified Hagerman and Kinnefors procedure was developed for measurements with noise when the slope is steeper. A new modification can perhaps be made in order to develop it for the purpose of SRT measurement, but at this stage it is not a feasible

method for this measurement. The original Hagerman and Kinnefors procedures were developed for use with five-word sentences.

3. The Brand and Kollmeier procedures were also developed for use with five-word sentences. The procedures can be modified, but they need to be administered by a computer and are not feasible for this use now.
4. The adaptive stimuli method with 30-words sets (10 utterances) at 5 dB intervals with a high level of 80 % and a low level of 20 % may be feasible for these measurements. The threshold is estimated with the curve-fitting method. Section 5.4.7.2 gives the standard deviation of 0.74 dB for hypothetical subject HS1 which deteriorates to 2.2-2.4 dB for the other hypothetical subjects. The mean number of tested items was 154 for HS1 with a standard deviation of 13.
5. The quick-speed test based on the constant stimuli method used with three-word sets (one utterance) at 1.5 dB intervals with 30 utterances in total is another feasible option for SRT measurement. The threshold is estimated by the counting method. Section 5.4.7.3 gives the standard deviation as 1.1 dB for HS1, increasing to 1.9-2.1 dB for HS2-HS4. A maximum of 90 test items are used with this procedure.

For "HiST taleaudiometri" we decided to recommend both procedure 4 and procedure 5 from this list. They should both give SRTs with reasonable accuracy for comparison with the pure tone thresholds, and they are quick to administer. The learning effects will be partially compensated for because most of the words initially are recognizable. The selection of method can be made by the institution or the audiologist performing the test.

5.5.2 Suprathreshold measurements with monosyllabic words

Maximum speech recognition score (PB_{max}) is the most important parameter measured using the monosyllabic words. At loud levels the intelligibility sometimes decreases (PB_{min}). Another parameter it may be interesting to estimate is the rollover index ($(PB_{max}-PB_{min})/PB_{max}$). An alternative to measuring these parameters is to draw the performance-intensity curve, which also can be used to estimate the parameters. In Norway the common approach among audiologists has been to draw the performance-intensity curve when doing this type of measurements. Scores obtained using 10-word sets measured at 5 or 10 dB intervals have provided the basis for the curve. Internationally, it is customary to measure maximum speech recognition score using a 50-word list at the level expected to give maximum recognition score. This level may have been estimated from the SRT or measured through initial testing procedures using 10-word sets.

In the simulations hypothetical subjects HS3 and HS4 have a maximum speech recognition score of 80 %. Equation 5.6 gives a standard deviation of 5.7 % from the binomial distribution when measuring scores with 50 words for this situation. In "HiST taleaudiometri" (Øygarden 2009, pp. 26-27) simulated measurements of the performance-intensity curve are made using 10-word sets at 5 dB intervals. The estimated maximum speech recognition score has a mean value of 83 % and the standard deviation is 6.5 % for HS3. For HS4 the mean value is 79 % and the standard deviation 7.5 %. The correct rollover index should be 0 for HS3 and 0.5 for HS4. The simulations give a mean value of 0.07 and a standard deviation of 0.1 for HS3, and a mean value of 0.48 and a standard deviation of 0.2 for HS4. An extra simulation for HS4 was performed in order to simulate measurement of rollover index with 50-word tests for PB_{max} and PB_{min}. This gave the results of 0.50 for the mean value of the rollover index and 0.095 for the standard deviation.

The results of these simulations show that the estimated maximum speech recognition score scores will be subject to somewhat greater variability when the measuring is performed using 10-word sets at 5 dB intervals and drawing the performance-intensity curve than if the measuring is performed directly using 50-word sets. However, the difference is not a large one, and we therefore decided that we can continue to measure maximum speech recognition score with 10-word sets if we measure at 5 dB intervals and draw the performance-intensity curve visually interpolated from the scores.

The estimated rollover index had twice the standard deviation with the curve-fitting routine compared to the simulated measuring of PB_{max} and PB_{min} using 50-word sets. Our recommendation is that if measurement of the rollover index is needed, this should be estimated by means of measurements of PB_{max} and PB_{min} using 50 word sets.

5.5.3 Signal-to-noise ratio measurements.

Both three-word utterances and five-word sentences are available in "HiST taleaudiometri" for measurements of signal-to-noise ratios. The steep slopes of 16 %/dB for three-word utterances and 14 %/dB for five-word sentences are crucial for obtaining measurements with good accuracy. It has been shown that the procedures selected for detecting differences in section 5.4.8.2 are the best alternatives for measuring signal-to-noise ratios because these were the procedures with the smallest standard deviations. The procedures for five-word sentences presented in Tables 5.6 and 5.7 can be ranked according to their standard deviations. The procedures with the smallest deviations top the list:

1. The Hagerman and Kinnefors S/N-threshold method, modified for use with 25 sentences. The standard deviation is 0.343.
2. The Brand and Kollmeier A1 threshold method, modified for use with 30 sentences. The standard deviation is 0.343.
3. The Hagerman and Kinnefors SRT-threshold method, modified for use with 25 sentences. The standard deviation is 0.382.
4. The Brand and Kollmeier A2 threshold and slope method, modified for use with 30 sentences. The standard deviation is 0.395.
5. The constant stimuli method with the counting routine, 5 word sets, modified to measure 30 levels over a 20 dB range, 150 words/session. The standard deviation is 0.464.
6. The 80-20 % adaptive stimuli method used with the curve fitting routine, 50-word sets, modified to a starting level 5 dB over the threshold measured at 5 dB intervals. The standard deviation is 0.502.
7. The constant stimuli method used with the counting routine, 5 word sets at 0.75 dB intervals, 150 words/session. The standard deviation is 0.537.
8. ISO 8253-3 Threshold level procedure A, modified to a starting level of 5 dB over the threshold, 50 words measured at 4 dB intervals. The standard deviation is 0.542.
9. The 80-20 % adaptive stimuli method with the curve-fitting routine, 50-word sets at 10 dB intervals with a starting level of 15 dB over the threshold. The mean value of words/session is 153 (from Table 5.6). The standard deviation is 0.808.
10. The constant stimuli method used with the counting routine, 5-word sets at 2.5 dB intervals, 50 words/session (from Table 5.6). The standard deviation is 0.930.

Even though the procedures of Hagerman and Kinnefors (1 and 3) were developed for the Swedish Hagerman sentences with different slopes they give a very good performance here when they are modified for use with more sentences than originally proposed. When high accuracy measurement is required these methods need to be considered. Similarly, the procedures of Brand and Kollmeier (2 and 4) also perform very well when used in combination with 150 words/session. The Brand and Kollmeier procedures are dependent on a computer administering the test, which is not realized in this release of “HiST taleaudiometri”, but may be considered at a later stage. Procedure 5 shows promising results but is not realized in “HiST taleaudiometri” yet. Procedure 8 might be an alternative, but since it uses a non-standard 4 dB interval it is not among the recommended procedures in "HiST taleaudiometri".

The recommended procedures in "HiST taleaudiometri" for use with five-word sentences are numbers 6, 7, 9 and 10. For quick measurements with moderate accuracy requirements procedures 9 and 10 are easy to administer, and procedure 10 is the fastest. Procedure 10 is the one used in the tests described in sections 6.1.3.6, 6.1.3.7 and 6.1.3.12. If better accuracy is required procedures 6 and 7 can be used, resulting in a small increase in the time required to perform the measurements. Procedure 7 is used in the test described in section 6.1.2.13.

The procedures for three-word utterances in Tables 5.6 and 5.7 can be ranked according to their standard deviations. The procedures with the smallest deviations top the list:

1. New method: Modified Hagerman and Kinnefors threshold method for pairs of three-word utterances. Level change rule -3, -2, -1, 0, 1, 2 and 3 dB for scores 6, 5, 4, 3, 2, 1 and 0. 22 pairs in measurement. The standard deviation is 0.292.
2. The 80-20 % adaptive stimuli method used with the curve-fitting routine, 30-word sets modified to a starting level of 5 dB over threshold measured at 2 dB intervals. The standard deviation is 0.401.
3. The constant stimuli method used with the counting routine, 3-word sets modified to measure 50 levels over a 20 dB range, 150 words/session. The standard deviation is 0.464.
4. ISO 8253-3 Threshold level procedure A modified to a starting level of 5 dB over the threshold, 30 words in set, 2 dB intervals (from Table 5.6). The standard deviation is 0.51.
5. The 80-20 % adaptive stimuli method used with the curve-fitting routine, 30-word sets measured at 5 dB intervals (from Table 5.6). The standard deviation is 0.552.
6. The constant stimuli method used with the counting routine, 3-word sets measured at 1.5 dB intervals, 90 words/session (from Table 5.6). The standard deviation is 0.837.

If measurements with very good accuracy are required, procedure 1 represents a good alternative. Procedure 2 is easy to implement and also gives good accuracy. Procedure 3 is not used in "HiST taleaudiometri". There is no reason to recommend procedure 4 as procedure 2 gives better results and is no more complicated to use.

The recommended procedures in "HiST taleaudiometri" for use with three-word utterances are numbers 5 and 6. The results produced by procedures 1-4 in the list are somewhat better, but procedures 5 and 6 are faster.

5.6 Conclusion

In this chapter simulations used to evaluate several different procedures for measuring speech audiometry scores were discussed. Some of these results are included in the report “HiST taleaudiometri” (Øygarden 2009) in order to allow users of the tests to evaluate the performance of the different methods when selecting which test to use for a specific situation.

In section 5.5 we selected which of these tests to recommend as part of the “HiST taleaudiometri” report on the basis of the simulations. The selected tests are time-efficient. Recommendations for tests with better accuracy are also given in section 5.5.

Chapter 6

Applications and recommendations

6.1 “HiST taleaudiometri”

The culmination of all the recording of speech material, manipulation of wave files, testing with or without noise and simulations of speech audiometry measurements is the speech audiometry set “HiST taleaudiometri” (Øygarden, 2009), made available by Sør-Trøndelag University College (HiST is short for the Norwegian of the college, “Høgskolen i Sør-Trøndelag”; and “taleaudiometri” is the Norwegian word for speech audiometry). The speech audiometry set consists of a report, two CDs and one audio DVD. Section 6.1.3 below gives a description of the material made available on the disks belonging to the set.

The report included in the “HiST taleaudiometri” set starts with a short introduction of the elements making up the speech audiometry set. This is followed by a detailed description of each type of test including the list of words used, their deployment and simulations in order to provide the reader with an impressions of the accuracy of the procedures. Then the results are given for the sound pressures levels of the words and sentences measured with the earphones on a coupler. Finally, measurement form originals for copying are provided.

6.1.1 Introduction

The following raw data was used to generate the material included in ”HiST taleaudiometri”:

- Five-word sentences. The development of these was described in Chapter 2. 400 wave files are available, making it possible to realize the 10 000 lists containing all of the 100 000 sentences available with this method. A nomenclature was developed, making it possible to control the realized lists and sentences in order to avoid repetition. 80 of the 10 000 lists possible are realized in "HiST taleaudiometri" at this stage. The nomenclature is presented in Appendix F.

- Three-word utterances. These were developed from the five-word sentences as described in Chapter 3. 200 wave files are available, making it possible to realize 100 lists containing all of the 1000 utterances available. 47 of the 100 lists possible are realized in "HiST taleaudiometri" at this stage. The same nomenclature that is used for five-word sentences is used here to keep control of lists and sentences. The nomenclature is presented in Appendix F.

- Monosyllabic words. 160 monosyllabic words were used to generate lists as described in Chapter 4.

- Monosyllabic numerals (digit triplets). These are taken from the old Norwegian speech audiometry set made by Sverre Quist-Hanssen. The triplets were cut out from the CD, and a certain amount of digital noise reduction was performed before they were spliced together in the original order with a standardised pause between the triplets. The level was adjusted in order to make it uniform with the other tests in "HiST taleaudiometri". The speech recognition threshold was measured to -3.4 dB HL and the slope to 17 %/dB for young normal-hearing subjects for monosyllabic numerals.

- Speech noise. The generation of our speech noise was described in section 2.2.2.6. When speech audiometry sentences/utterances were produced with speech noise on the other channel of the CD-disk, the noise was cut out from a random position in a 30 seconds long wave file containing the noise.

Section 6.1.3 below describes what type of test material is realized in "HiST taleaudiometri" on the basis of this raw material.

6.1.2 Calibration of "HiST taleaudiometri"

The three-word utterances mentioned above were selected as the standard material for measurement of the speech recognition threshold without noise in "HiST taleaudiometri". The decision was made to design the calibration signal such that the speech recognition threshold for three-word utterances was 0 dB HL. The previous Norwegian speech audiometry developed by

Table 6.1 Speech recognition thresholds and slopes for the different materials in "HiST taleaudiometri" in silence and with noise.

	SRT L_{eq} [dBC]	Standard deviation [dB]	SRT [dB "HL"]	Slope [%/dB]	SRT in noise [dB SNR]	Slope in noise [%/dB]
Digit triplets	19.2	1.4	-3.4	17		
Three-word utterances (reference)	22.6	1.0	0.0	10	-6.2	16
Five-word sentences	23.6	0.8	1.0	10	-6.0	14
Monosyllabic words	29.1	2.1	6.5	7		

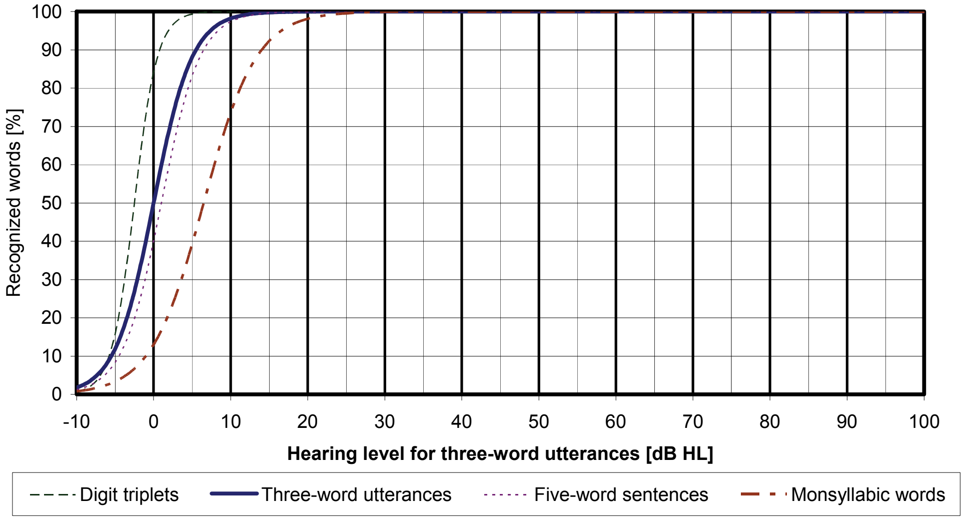


Figure 6.1 Normal performance-intensity curves for digit triplets, three-word utterances, five-word sentences and monosyllabic words.

Quist-Hanssen had selected the speech recognition threshold for monosyllabic PB words as the reference threshold and equalized all the word lists (monosyllabic PB words, spondees and digit triplets) to a uniform intelligibility threshold (Quist-Hanssen 1966). We decided not to continue this tradition but to release all the material at the same level on the CDs. The speech recognition thresholds for the different materials in "HiST taleaudiometri" were measured in the first laboratory test (results in sections 2.3.4, 3.3.2 and 4.3.2) and are presented in Table 6.1 together with the measured C-weighted sound pressure levels of the thresholds. The sound pressure levels are equivalent levels with the silent parts between speech sounds removed, measured from the supraaural earphones TDH 39 with MX-41/AR cushions mounted on an IEC 60318-3 (1998) acoustic coupler. The slopes of the performance-intensity curves are also presented. For three-word utterances and five-word sentences the thresholds expressed in dB SNR and the slopes are also presented for measurements in noise (sections 2.3.5 and 3.3.3). Chapter 20 in "HiST taleaudiometri" (Øygarden 2009) presents the levels of all the words, utterances, sentences and lists contained on the CDs.

The normal performance-intensity curves for the material included in the "HiST taleaudiometri" set are presented in Figure 6.1.

The thresholds for the Quist-Hanssen material were obtained in the same session in which the thresholds for "HiST taleaudiometri" were measured. The speech recognition thresholds for the Quist-Hanssen material were -5.0 dB HL for the digit triplets and monosyllabic words, and -2.0 dB HL for the spondees .

6.1.3 The contents of the ”HiST Taleaudiometri” set

6.1.3.1 CD1 tracks 1-10, monosyllabic words

- Track 1: VU-adjustment tone
- Tracks 2-10: Lists 1-9, each list containing 50 monosyllabic words. 4-second interval between the start point of each word

The development of this material was presented in Chapter 4.

6.1.3.1.1 Deployment

Two deployment strategies are proposed:

Either to use a complete 50-word list to measure maximum speech recognition score at a suitable level.

Or to measure the performance-intensity curve using groups of 10 words at 5 dB intervals. Choose a starting level of PTA + 25 dB. Change levels in 5 dB intervals up/down to register both 100 % and 0 % speech recognition. To obtain maximum speech recognition score measure upwards in 5 dB intervals until 85 dB HL or discomfort is reached.

6.1.3.2 CD1 tracks 11-13, words for children selected by Rikshospitalet.

- Tracks 11-13: RC1-3, each list containing 50 monosyllabic words selected by Rikshospitalet University Hospital for the purpose of testing children

The material was presented in section 4.2.2.2.

6.1.3.2.1 Deployment

We recommend that the deployment is adjusted to the age of the children. RC1-RC2 are for the youngest children, while RC3 contains some more advanced words for older children.

6.1.3.3 CD1 tracks 14-34, three-word utterances for speech recognition threshold measurements

- Track 14: VU-adjustment tone
- Tracks 15-34: 20 lists, each list containing 10 three-word utterances. 5-second pauses between utterances. Speech on left channel. Noise (optional to use) is recorded on the right channel, starting 0.5 seconds before speech and ending 0.5 seconds after speech.

The development of this material was presented in Chapter 3.

6.1.3.3.1 Deployment

We recommend the use of the adaptive stimuli method with curve fitting (section 5.3.6.2), starting level PTA + 15 dB, first getting a high score over 80 % and measuring in 5 dB intervals down to a low score below 20 %. The score is calculated from the proportion of correctly recognized words in the 30-word list. The threshold is the level giving 50 % score on the fitted curve.

6.1.3.4 CD1 tracks 35-40, quick-speed test, three-word utterances

- Track 35: VU-adjustment tone

- Tracks 36-40: Four quick-speed test lists. The test lists consist of 30 three-word utterances where the level is reduced by 1.5 dB for each utterance. The threshold is estimated by the counting method. 4.75-second pauses between utterances. Speech on left channel. Noise (optional to use) is recorded on the right channel, starting 0.5 seconds before speech and ending 0.25 seconds after speech.

6.1.3.4.1 Deployment

The threshold is estimated by the counting method; all the recognized words on the list are counted. Two alternative procedures are proposed:

Formula-based. Choose a starting level of 25-30 dB higher than the PTA. Play one list and count the number of words. If the test subject has no prior experience with these lists it is important that almost all of the words in the 10 first sentences are recognized as training for the rest of the test. The measuring can be discontinued if no recognition is accomplished for four consecutive sentences. The threshold can be calculated by:

$$\text{Starting level} - \frac{\text{number of recognized words}}{2} \quad [\text{dB HL}]$$

Table-based. A table where the test administrator uses the PTA as input to find a recommended starting level is presented in “HiST taleaudiometri” (Chapter 8). After testing the table will enable him or her to find the threshold from the number of words recognized.

6.1.3.5 CD2 tracks 1-11, five-word sentences for speech recognition threshold measurements

- Track 1: VU-adjustment tone
- Tracks 2-11: 10 lists, each containing 10 five-word sentences. 7-second pauses between sentences. Speech on left channel. Noise (optional to use) is recorded on the right channel, starting 0.5 seconds before speech and ending 0.5 seconds after speech.

6.1.3.5.1 Deployment

We recommend that the adaptive stimuli method (section 5.3.6.2) is used, with a starting level of PTA + 15 dB, first getting a high score over 80 % and measuring in 10 dB intervals down to a low score below 20 %. The score is calculated from the proportion of correctly recognized words in the 50-word list. The threshold is the level giving 50 % score on the fitted curve.

6.1.3.6 CD2 tracks 12-16, quick-speed test, five-word sentences

- Track 12: VU-adjustment tone
- Tracks 13-16: Four quick-speed test lists. The test lists consist of 20 five-word sentences where the level is reduced by 2.5 dB for each sentence. The threshold is estimated by the counting method. 6.75-second pauses between sentences. Speech on left channel. Noise (optional to use) is recorded on the right channel, starting 0.5 seconds before speech and ending 0.25 seconds after speech.

6.1.3.6.1 Deployment

The threshold is estimated by the counting method; all the recognized words on the list are counted. Two alternative procedures are proposed:

Formula-based. Choose a starting level of 25-30 dB higher than the PTA. Play one list and count the number of words. If the test subject has no prior experience with these lists it is important that almost all of the words in the 10 first sentences are recognized as training for the rest of the test. The measuring can be stopped if no recognition is accomplished for four consecutive sentences. The threshold can be calculated by:

$$\text{Starting level} - \frac{\text{number of recognized words}}{2} \quad [\text{dB HL}]$$

Table-based. A table where the test administrator uses the PTA as input to find a recommended starting level is presented in “HiST taleaudiometri” (Chapter 10). After testing the table will enable him or her to find the threshold from the number of words recognized.

6.1.3.7 CD2 tracks 17-36, binaural tests with earphones

- Tracks 17 and 27: VU-adjustment tone
- Tracks 18-28 and 28-36: two equally configured sets of binaural tests. Each set consists of 9 lists, one list on each track. Each list consists of 10 five-word sentences with speech designed by the quick-speed test method where the level is reduced by 2.5 dB for each sentence. There is a different speech and noise setup for each list.: 1 - Speech binaural, no noise. 2 - Speech binaural, noise on left ear. 3 - Speech binaural, noise on right ear. 4 - Speech binaural phase shifted, noise binaural. 5 - Speech binaural, noise binaural phase shifted. 6 - Speech binaural, noise temporally simulated in left ear by delaying noise by 0.6 ms in right ear. 7 - binaural, noise

temporally simulated in right ear by delaying noise by 0.6 ms in left ear. 8 - Speech binaural, noise binaural uncorrelated. 9 - Speech binaural, noise binaural.

6.1.3.7.1 Deployment

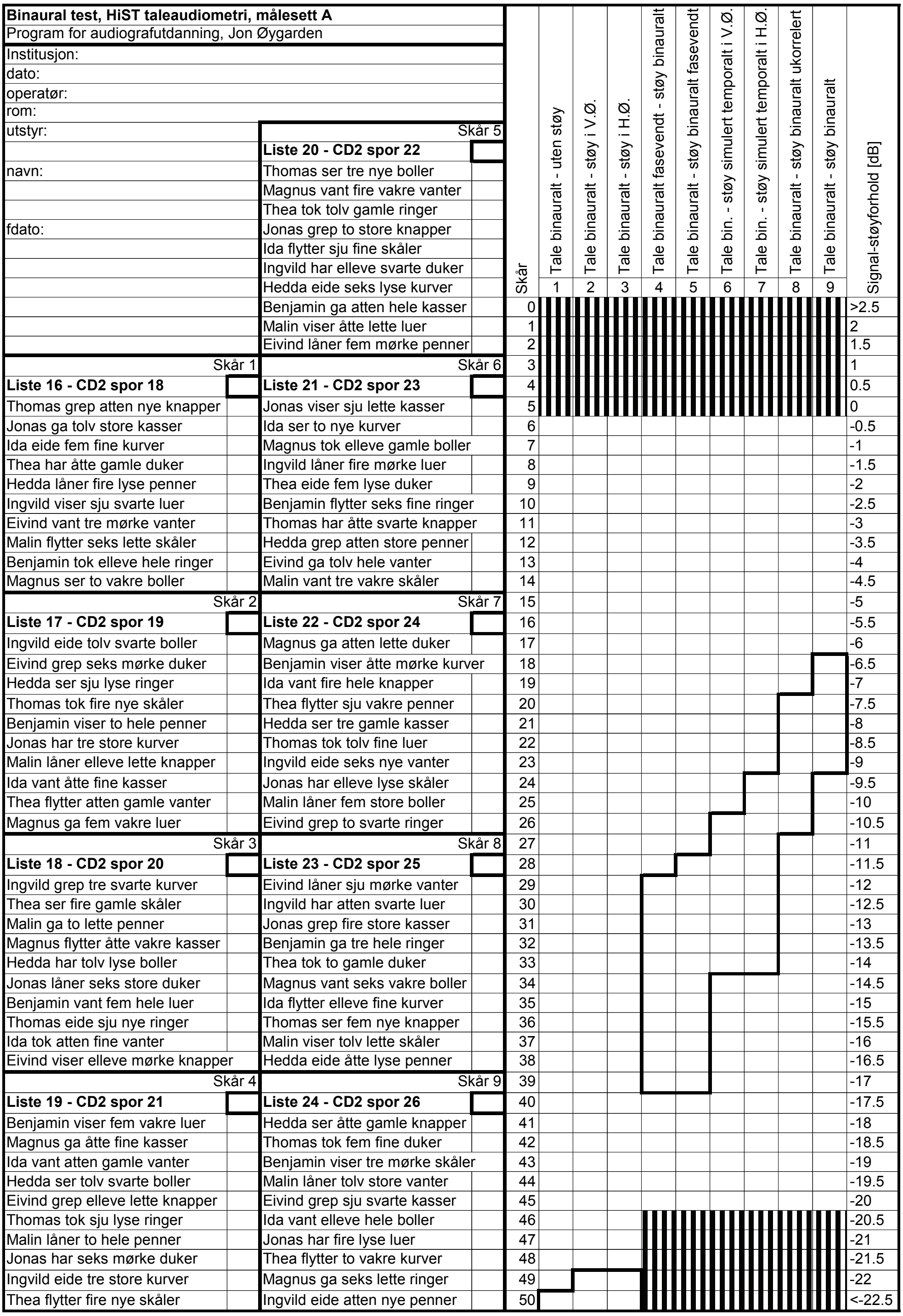
A test form for these measurements was developed for each of these two test sets. An example of the test form for test set A is shown in Figure 6.2. Before binaural tests can be performed with earphones on an audiometer, the phasing of the earphones must be checked. A special test signal has been prepared on CD2 track 63 for this purpose. If the earphones are not in phase, the phase has to be changed on one earphone.

The proposed starting level is 45 dB over the speech recognition level. Play track 18 or 28 and check with the test subject whether the level is comfortable; adjust if necessary. For this track and the 8 consecutive tracks, count the number of words recognized in the 10 sentences and register on the test form. For the first three tracks normal-hearing persons will achieve almost full score because the tracks are intended as training tracks allowing the test subjects to get acquainted with the test material and the testing situation.

Use the scores for the 10 sentences of each track to put an X in the corresponding column on the right side of the test form. The signal-to-noise ratios for the speech recognition threshold of each speech/noise situation can then be found. The thicker lines in the form highlight the threshold ± one standard deviation from a small group of young normal-hearing test persons.

The protocol will show a binaural profile when finished and may be of help when counselling persons who have problems with directional hearing. Many of the psychoacoustic subtests selected here are used in a masking level difference (MLD) setup.

These measurement sets were developed in the hope that further studies may be performed in order to assess the value of these measurements for different groups of people with hearing problems.



Binaural test, HiST taleaudiometri, målesett A

Program for audiografutdanning, Jon Øygarden

Institusjon:

dato:

operatør:

rom:

utstyr:

navn:

fdato:

Skår 1

Liste 16 - CD2 spor 18 □

- Thomas grep atten nye knapper
- Jonas ga tolv store kasser
- Ida eide fem fine kurver
- Thea har åtte gamle duker
- Hedda låner fire lyse penner
- Ingvild viser sju svarte luer
- Eivind vant tre mørke vanter
- Malin flytter seks lette skåler
- Benjamin tok elleve hele ringer
- Magnus ser to vakre boller

Skår 2

Liste 17 - CD2 spor 19 □

- Ingvild eide tolv svarte boller
- Eivind grep seks mørke duker
- Hedda ser sju lyse ringer
- Thomas tok fire nye skåler
- Benjamin viser to hele penner
- Jonas har tre store kurver
- Malin låner elleve lette knapper
- Ida vant åtte fine kasser
- Thea flytter atten gamle vanter
- Magnus ga fem vakre luer

Skår 3

Liste 18 - CD2 spor 20 □

- Ingvild grep tre svarte kurver
- Thea ser fire gamle skåler
- Malin ga to lette penner
- Magnus flytter åtte vakre kasser
- Hedda har tolv lyse boller
- Jonas låner seks store duker
- Benjamin vant fem hele luer
- Thomas eide sju nye ringer
- Ida tok atten fine vanter
- Eivind viser elleve mørke knapper

Skår 4

Liste 19 - CD2 spor 21 □

- Benjamin viser fem vakre luer
- Magnus ga åtte fine kasser
- Ida vant atten gamle vanter
- Hedda ser tolv svarte boller
- Eivind grep elleve lette knapper
- Thomas tok sju lyse ringer
- Malin låner to hele penner
- Jonas har seks mørke duker
- Ingvild eide tre store kurver
- Thea flytter fire nye skåler

Skår 5

Liste 20 - CD2 spor 22 □

- Thomas ser tre nye boller
- Magnus vant fire vakre vanter
- Thea tok tolv gamle ringer
- Jonas grep to store knapper
- Ida flytter sju fine skåler
- Ingvild har elleve svarte duker
- Hedda eide seks lyse kurver
- Benjamin ga atten hele kasser
- Malin viser åtte lette luer
- Eivind låner fem mørke penner

Skår 6

Liste 21 - CD2 spor 23 □

- Jonas viser sju lette kasser
- Ida ser to nye kurver
- Magnus tok elleve gamle boller
- Ingvild låner fire mørke luer
- Thea eide fem lyse duker
- Benjamin flytter seks fine ringer
- Thomas har åtte svarte knapper
- Hedda grep atten store penner
- Eivind ga tolv hele vanter
- Malin vant tre vakre skåler

Skår 7

Liste 22 - CD2 spor 24 □

- Magnus ga atten lette duker
- Benjamin viser åtte mørke kurver
- Ida vant fire hele knapper
- Thea flytter sju vakre penner
- Hedda ser tre gamle kasser
- Thomas tok tolv fine luer
- Ingvild eide seks nye vanter
- Jonas har elleve lyse skåler
- Malin låner fem store boller
- Eivind grep to svarte ringer

Skår 8

Liste 23 - CD2 spor 25 □

- Eivind låner sju mørke vanter
- Ingvild har atten svarte luer
- Jonas grep fire store kasser
- Benjamin ga tre hele ringer
- Thea tok to gamle duker
- Magnus vant seks vakre boller
- Ida flytter elleve fine kurver
- Thomas ser fem nye knapper
- Malin viser tolv lette skåler
- Hedda eide åtte lyse penner

Skår 9

Liste 24 - CD2 spor 26 □

- Hedda ser åtte gamle knapper
- Thomas tok fem fine duker
- Benjamin viser tre mørke skåler
- Malin låner tolv store vanter
- Eivind grep sju svarte kasser
- Ida vant elleve hele boller
- Jonas har fire lyse luer
- Thea flytter to vakre kurver
- Magnus ga seks lette ringer
- Ingvild eide atten nye penner

Skår	Tale binauralt - uten støy	Tale binauralt - støy i V.Ø.	Tale binauralt - støy i H.Ø.	Tale binauralt fasevendt - støy binauralt	Tale binauralt - støy binauralt fasevendt	Tale bin. - støy simulert temporalt i V.Ø.	Tale bin. - støy simulert temporalt i H.Ø.	Tale binauralt - støy binauralt ukorrelert	Tale binauralt - støy binauralt	Signal-støyforhold [dB]
	1	2	3	4	5	6	7	8	9	
0	▨	▨	▨	▨	▨	▨	▨	▨	▨	>2.5
1	▨	▨	▨	▨	▨	▨	▨	▨	▨	2
2	▨	▨	▨	▨	▨	▨	▨	▨	▨	1.5
3	▨	▨	▨	▨	▨	▨	▨	▨	▨	1
4	▨	▨	▨	▨	▨	▨	▨	▨	▨	0.5
5	▨	▨	▨	▨	▨	▨	▨	▨	▨	0
6										-0.5
7										-1
8										-1.5
9										-2
10										-2.5
11										-3
12										-3.5
13										-4
14										-4.5
15										-5
16										-5.5
17										-6
18										-6.5
19										-7
20										-7.5
21										-8
22										-8.5
23										-9
24										-9.5
25										-10
26										-10.5
27										-11
28										-11.5
29										-12
30										-12.5
31										-13
32										-13.5
33										-14
34										-14.5
35										-15
36										-15.5
37										-16
38										-16.5
39										-17
40										-17.5
41										-18
42										-18.5
43										-19
44										-19.5
45										-20
46				▨	▨	▨	▨	▨	▨	-20.5
47				▨	▨	▨	▨	▨	▨	-21
48				▨	▨	▨	▨	▨	▨	-21.5
49				▨	▨	▨	▨	▨	▨	-22
50				▨	▨	▨	▨	▨	▨	<-22.5

Figure 6.2 The form for binaural test set A.

6.1.3.8 CD2 tracks 37-38, monosyllabic numerals (digit triplets).

- Track 37: VU-adjustment tone
- Track 38: 60 groups of three monosyllabic numerals (digit triplets). 4-second pauses between the digit triplets.

6.1.3.8.1 Deployment

This test is included for use on test subjects who have problems with more complex speech material. Measure the performance-intensity curve using groups of 9 numbers (three triplets) at 5 dB intervals. Choose a starting level of PTA + 5 dB. Change levels in 5 dB intervals up/down to register both 100 % and 0 % speech recognition.

6.1.3.9 CD2 tracks 39-63, signals for calibration

The levels given in this section are equivalent levels relative to the 16 bit dynamic range on CDs. dBFSSQ means dB Full Scale for SQuare waves. A full range square wave will have a level of 0 dBFSSQ.

- Track 39: 1000 Hz, 1/3-octave noise at -28.5 dBFSSQ, 1 minute length. For calibration.
- Tracks 40-58: 125-8000 Hz, 1/3 octave noise at -28.5 dBFSSQ, 15 seconds length. For checking frequency response.
- Tracks 59-61: 250, 500 and 1000 Hz sinus tone at -6 dBFSSQ, 1 minute length. For checking distortion.
- Track 62: 1000 Hz sinus tone at -8.5 dBFSSQ, 10 seconds length. For preliminary calibration if audiometer is calibrated for Quist-Hanssen speech audiometry. (The same signal can also be found on CD1 track 41)
- Track 63: Binaural phase test signal.

6.1.3.10 DVD title 1, three-word utterances for free field audiometry

- Chapters 1-4: Four quick-speed test lists. The test lists consist of 30 three-word utterances where the level is reduced by 2 dB for each utterance. The threshold is estimated by the counting method. 4-second pauses between utterances. Speech on centre channel, without noise. The starting level is locked to 65 dB sound pressure level on the DVD tests.

6.1.3.10.1 Deployment

For measurement of speech recognition threshold in free field. The threshold is estimated by the counting method; all the recognized words on the list are counted. Two alternative procedures are proposed:

Formula-based. Play one list and count the number of recognized words. If the test subject has no prior experience with these lists it is important that almost all of the words in the 10 first utterances are recognized as training for the rest of the test. The measuring can be discontinued if no recognition is accomplished for four consecutive utterances. The threshold expressed in dB SPL can be calculated by:

$$65 - 2\,\frac{\text{number of recognized words}}{3} \quad [\text{dB SPL}]$$

Table-based. A table where the test administrator can find the threshold from the number of words recognized is included in "HiST taleaudiometri" (Chapter 15).

6.1.3.11 DVD title 1, five-word sentences for free field audiometry

- Chapters 5-8: Four quick-speed test lists. The test lists consist of 20 five-word sentences where the level is reduced by 3 dB for each sentence. The threshold is estimated by the counting method. 6-second pauses between utterances. Speech on centre channel, without noise. The starting level is locked to 65 dB sound pressure level on the DVD tests.

6.1.3.11.1 Deployment

For measurement of speech recognition threshold in free field. The threshold is estimated by the counting method; all the recognized words on the list are counted. Two alternative procedures are proposed:

Formula-based. Play one list and count the number of recognized words. If the test subject has no prior experience with these lists it is important that almost all of the words in the 10 first sentences are recognized as training for the rest of the test. The measuring can be discontinued if no recognition is accomplished for four consecutive sentences. The threshold expressed in dB SPL can be calculated by:

$$65 - 3\,\frac{\text{number of recognized words}}{5}\quad [\text{dB SPL}]$$

Table-based. A table where the test administrator can finds the threshold from the number of words recognized is included in “HiST taleaudiometri” (Chapter 15).

6.1.3.12 DVD title 2, 5 five-word sentences in noise and reverberation

Four equally configured sets have been prepared; each title contains one set which consists of 6 chapters containing 10 five-word sentences of speech designed by the quick-speed test method; the level is reduced by 2.5 dB for each sentence. There is a different speech, noise and reverberation setup for each chapter. The speech is in the centre channel; uncorrelated noise in four surround channels. The noise starts 1 second before and ends 0.25 seconds after the speech. On some chapters reverberation is simulated, with 1.5 seconds reverberation time.

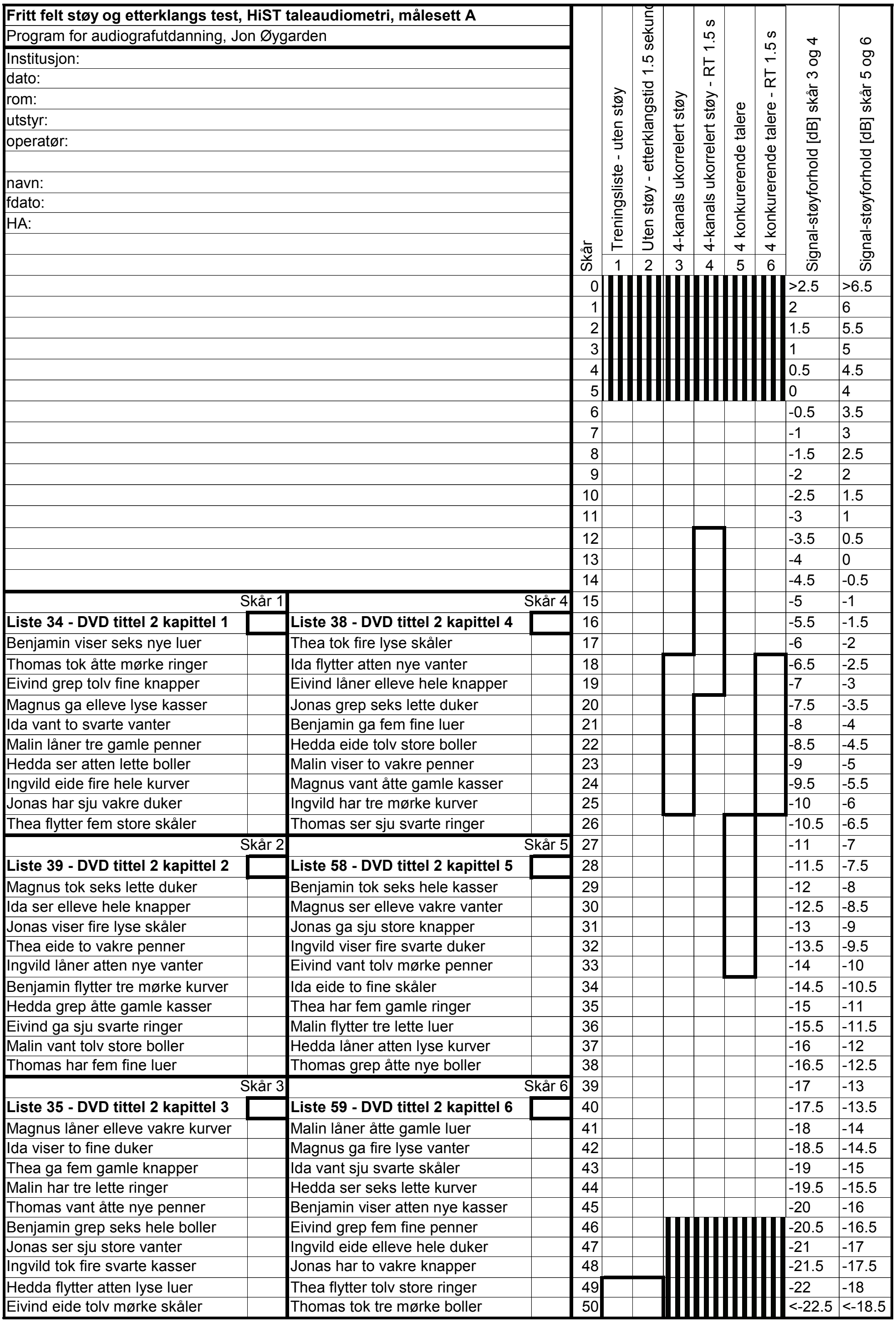
- Chapter 1: Training list without noise and reverberation.
- Chapter 2: Speech + reverberation.
- Chapter 3: Speech + 4 channels of uncorrelated noise.
- Chapter 4: Speech + 4 channels of uncorrelated noise + reverberation.
- Chapter 5: Speech + 4 concurrent speakers (two women and two men). One speaker in each of the surround channels. The level of the concurrent speakers is reduced by 10 dB compared to the centre channel.
- Chapter 6: Like chapter 5 but with reverberation.

6.1.3.12.1 Deployment

A test form for these measurements was developed for each of the test sets. An example of the test form for test set A is shown in Figure 6.3.

The starting levels are locked to 65 dB SPL on the DVD. Play the first chapter; for this chapter and the 5 consecutive chapters count the number of words recognized in 10 sentences and register on the test form. For the first two chapters normal-hearing persons will achieve almost full score because the chapters are intended as training, allowing the test subjects to get acquainted with the test material and the testing situation.

Use the scores for the 10 sentences of each chapter to put an X in the corresponding column on the right side of the test form. The signal-to-noise



Fritt felt støy og etterklangs test, HiST taleaudiometri, målesett A

Program for audiografutdanning, Jon Øygarden

Institusjon:
dato:
rom:
utstyr:
operatør:

navn:
fdato:
HA:

	Skår 1		Skår 4
Liste 34 - DVD tittel 2 kapittel 1	□	**Liste 38 - DVD tittel 2 kapittel 4**	□
Benjamin viser seks nye luer		Thea tok fire lyse skåler	
Thomas tok åtte mørke ringer		Ida flytter atten nye vanter	
Eivind grep tolv fine knapper		Eivind låner elleve hele knapper	
Magnus ga elleve lyse kasser		Jonas grep seks lette duker	
Ida vant to svarte vanter		Benjamin ga fem fine luer	
Malin låner tre gamle penner		Hedda eide tolv store boller	
Hedda ser atten lette boller		Malin viser to vakre penner	
Ingvild eide fire hele kurver		Magnus vant åtte gamle kasser	
Jonas har sju vakre duker		Ingvild har tre mørke kurver	
Thea flytter fem store skåler		Thomas ser sju svarte ringer	
	Skår 2		Skår 5
Liste 39 - DVD tittel 2 kapittel 2	□	**Liste 58 - DVD tittel 2 kapittel 5**	□
Magnus tok seks lette duker		Benjamin tok seks hele kasser	
Ida ser elleve hele knapper		Magnus ser elleve vakre vanter	
Jonas viser fire lyse skåler		Jonas ga sju store knapper	
Thea eide to vakre penner		Ingvild viser fire svarte duker	
Ingvild låner atten nye vanter		Eivind vant tolv mørke penner	
Benjamin flytter tre mørke kurver		Ida eide to fine skåler	
Hedda grep åtte gamle kasser		Thea har fem gamle ringer	
Eivind ga sju svarte ringer		Malin flytter tre lette luer	
Malin vant tolv store boller		Hedda låner atten lyse kurver	
Thomas har fem fine luer		Thomas grep åtte nye boller	
	Skår 3		Skår 6
Liste 35 - DVD tittel 2 kapittel 3	□	**Liste 59 - DVD tittel 2 kapittel 6**	□
Magnus låner elleve vakre kurver		Malin låner åtte gamle luer	
Ida viser to fine duker		Magnus ga fire lyse vanter	
Thea ga fem gamle knapper		Ida vant sju svarte skåler	
Malin har tre lette ringer		Hedda ser seks lette kurver	
Thomas vant åtte nye penner		Benjamin viser atten nye kasser	
Benjamin grep seks hele boller		Eivind grep fem fine penner	
Jonas ser sju store vanter		Ingvild eide elleve hele duker	
Ingvild tok fire svarte kasser		Jonas har to vakre knapper	
Hedda flytter atten lyse luer		Thea flytter tolv store ringer	
Eivind eide tolv mørke skåler		Thomas tok tre mørke boller	

	Treningsliste - uten støy	Uten støy - etterklangstid 1.5 sekund	4-kanals ukorrelert støy	4-kanals ukorrelert støy - RT 1.5 s	4 konkurerende talere	4 konkurerende talere - RT 1.5 s	Signal-støyforhold [dB] skår 3 og 4	Signal-støyforhold [dB] skår 5 og 6
Skår	1	2	3	4	5	6		
0	▨	▨	▨	▨	▨	▨	>2.5	>6.5
1	▨	▨	▨	▨	▨	▨	2	6
2	▨	▨	▨	▨	▨	▨	1.5	5.5
3	▨	▨	▨	▨	▨	▨	1	5
4	▨	▨	▨	▨	▨	▨	0.5	4.5
5	▨	▨	▨	▨	▨	▨	0	4
6							-0.5	3.5
7							-1	3
8							-1.5	2.5
9							-2	2
10							-2.5	1.5
11							-3	1
12							-3.5	0.5
13							-4	0
14							-4.5	-0.5
15							-5	-1
16							-5.5	-1.5
17							-6	-2
18							-6.5	-2.5
19							-7	-3
20							-7.5	-3.5
21							-8	-4
22							-8.5	-4.5
23							-9	-5
24							-9.5	-5.5
25							-10	-6
26							-10.5	-6.5
27							-11	-7
28							-11.5	-7.5
29							-12	-8
30							-12.5	-8.5
31							-13	-9
32							-13.5	-9.5
33							-14	-10
34							-14.5	-10.5
35							-15	-11
36							-15.5	-11.5
37							-16	-12
38							-16.5	-12.5
39							-17	-13
40							-17.5	-13.5
41							-18	-14
42							-18.5	-14.5
43							-19	-15
44							-19.5	-15.5
45							-20	-16
46			▨	▨	▨	▨	-20.5	-16.5
47			▨	▨	▨	▨	-21	-17
48			▨	▨	▨	▨	-21.5	-17.5
49			▨	▨	▨	▨	-22	-18
50			▨	▨	▨	▨	<-22.5	<-18.5

Figure 6.3 The form for free field test set A.

ratios for the speech recognition threshold of each speech/noise situation can then be found. The thicker lines in the form highlight the threshold ± one standard deviation from a small group of young normal-hearing test persons.

These measurement sets were developed in hope that further studies may be performed in order to assess the value of these measurements for different groups of people with hearing problems, for evaluating different hearing aids, different adjustments of hearing aids and/or for improvement over time with the use of hearing aids.

6.1.3.13 DVD title 6, five-word sentences for improved measurement accuracy

- Chapters 1-4: Four quick-speed test lists. The test lists consist of 30 five-word sentences where the level is reduced by 0.75 dB for each sentence. The threshold is estimated by the counting method. 7.25-second pauses between utterances. Speech on the centre channel, uncorrelated noise on the four surround channels. The noise starts 1 second before and ends 0.25 seconds after the speech. The starting level is locked to 65 dB sound pressure level on the DVD tests.

6.1.3.13.1 Deployment

The same type of measurement as on Chapter 3 in the preceding test, but with increased measurement accuracy because the intervals are reduced from 2.5 dB to 0.75 dB.

6.1.3.14 DVD title 7, Calibration sounds 1

- Chapters 1-5: 1000 Hz 1/3-octave noise in C, FL, FR, SL and SR channels for adjustment to 65 dB SPL.

6.1.3.15 DVD title 7, Calibration sounds 2

- Chapters 1-5: 125-8000 Hz, 1/3-octave noise in C, FL, FR, SL, SR channels; for adjustment to 65 dB SPL and checking of frequency response.
- Chapter 6-8: 250, 500 and 1000 Hz sinus tone at speech peak level, 1 minute length; for checking distortion.

- Chapter 9: 1000 Hz, 1/3-octave noise at centre channel, from 75 dB SPL to 0 dB SPL; every 5 seconds the noise is reduced by 5 dB. For checking linearity.

6.2 Recommendations for measurement of speech recognition threshold

The three-word utterances presented in Chapter 3 were developed for the purpose of measuring speech recognition thresholds. Several of the tests included in the "HiST taleaudiometri" set can be used for this purpose:

- Section 6.1.3.3 describes 20 lists consisting of 10 three-word utterances. The recommended measurement procedure is the adaptive stimuli method measured in 5 dB intervals with curve-fitting. The expected theoretical standard deviation for hypothetical subject HS1 is 0.74 dB, and this increases to 2.2-2.4 dB for the other hypothetical subjects. These tests can also be used with noise to measure signal-to-noise ratios.

- Section 6.1.3.4 describes four quick-speed test lists consisting of 30 three-word utterances. The threshold is estimated with the counting method. The expected theoretical standard deviation for hypothetical subject HS1 is 1.1 dB, and it rises to 1.9-2.1 dB for HS2-HS4. These tests can also be used with noise to measure signal-to-noise ratios.

- Section 6.1.3.10 describes four quick-speed test lists consisting of 30 three-word utterances. These lists are on the audio DVD-disk and are intended for free-field measurements of speech recognition thresholds with or without hearing aids. The expected theoretical standard deviation for hypothetical subject HS1 is 1.3 dB, and it increases to 1.9-2.4 dB for HS2-HS4.

6.3 Recommendations for measurement of maximum speech recognition score

The material containing the monosyllabic words presented in Chapter 4 was designed for measurement of maximum speech recognition score. Nine lists, each containing 50 monosyllabic words have been developed. Two alternatives for measuring maximum speech recognition score were proposed in section 6.1.3.1:

- Measure with a complete list of 50 words at the level expected to give maximum speech recognition score. When maximum speech recognition score is 80 % a theoretical standard deviation of 5.7 % is expected.

- Measure the performance-intensity curve using groups of 10 words at 5 dB intervals. When maximum speech recognition score is 80 % for hypothetical subjects HS3 and HS4 a theoretical standard deviation of 6.5-7.5 % is expected.

6.4 Recommendations for measurement of speech recognition for hearing aid evaluation

The five-word sentences and noise presented in Chapter 2 are recommended for measurement of speech recognition for hearing aid evaluation. "HiST taleaudiometri" contains several tests which can be used for this purpose:

- Section 6.1.3.5 presented 10 lists, each containing 10 five-word sentences. The recommended measurement procedure is the adaptive stimuli method deployed in 10 dB intervals with curve fitting. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.81 dB, and it increases to 2.1.-2.3 dB for hypothetical subjects HS2-HS4.

- Section 6.1.3.6 presented four quick-speed lists, each containing 20 sentences. The threshold is estimated using the counting method. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.93 dB, and this rises to 2.0-2.1 dB for HS2-HS4.

- Section 6.1.3.12 presents four test sets, each consisting of six test lists of speech in different noise and reverberation surroundings, available on the DVD-disk. The threshold is estimated using the counting method. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.93 dB, rising to 1.8-2.1 dB for HS2-HS4.

- Section 6.1.3.13 presents four quick-speed tests for improved measurement accuracy with four channel surround noise, available on the DVD-disk. The threshold is estimated using the counting method. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.54 dB, rising to 1.3-1.5 dB for HS2-HS4.

6.5 Recommendations for measurement of binaural performance

The method for generating five-words sentences with the possibility of making 100 000 different sentences gives good opportunities for realizing specialized tests. In order to obtain experience with the process of this deployment of the speech audiometry material we conducted some tests which can be evaluated at a later stage.

- Section 6.1.3.7 described two equally configured sets of binaural tests to be measured with earphones. Each set consists of 9 quick-speed lists with 10 sentences in each list. The threshold is estimated using the counting method. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.93 dB, rising to 1.8-2.1 dB for HS2-HS4

- This test was presented in section 6.4 for measurement of speech recognition in connection with hearing-aid evaluation, but it may also be used as a test of free-field binaural performance. Section 6.1.3.12 presented four test sets, each with six test lists containing speech in different noise and reverberation surroundings, available on the DVD. The threshold is estimated using the counting method. The expected theoretical standard deviation for a hypothetical subject with a slope of 14 %/dB is 0.93 dB, rising to 1.8-2.1 dB for HS2-HS4.

6.6 Further work

“HiST taleaudiometri” has been made available as a set for speech audiometry testing in Norway. It is intended as a replacement for the existing speech audiometry tests for Norwegian speakers. In Chapters 2-6 I described the methods used to select this speech audiometry material and gave recommendations for its deployment. It is my hope that the material will function well for the intended purpose. Nevertheless, some unanswered questions remain – questions to which the answers will hopefully be found by further work in relation to this material:

- Only young normal-hearing subjects were used for the normalizing tests in “HiST speech audiometry”. How will the material function when measuring is performed on hearing-impaired and/or elderly subjects? What relationship exists between other audiometry tests and the different speech audiometry tests included here?

- The nine lists of monosyllabic words are mixed to achieve good equivalence between the lists in theory; how this works in practice will have to be evaluated both on normal-hearing and hearing-impaired persons.

- The lists of monosyllabic words for children selected by Rikshospitalet have not been tested on children. Recommendations for use and evaluation of results need to be established.

- The quick-speed tests developed in “HiST taleaudiometri” have not been tried in a clinical situation. Do these tests amount to a fast and reliable method for measuring speech recognition thresholds on hearing-impaired subjects?

- The binaural test for earphones is realized in order to provide a method for evaluating certain psychoacoustic parameters connected with binaural performance. Are the normal limits chosen for this test correct? How does age and hearing loss influence the results? Is this test a help when counselling persons who report binaural hearing problems? Can the results of this test be compared with other methods of measuring binaural performance?

- The tests on the audio DVD, which include surround noise and reverberation, are intended both for compiling a profile for how well a subject performs in different noise situations and for making comparisons between different hearing aids. Is obtaining this profile for the individual helpful, and is there any relationship between this profile and the results of the binaural test for earphones? The test is quick to administer, but is its accuracy good enough for hearing aid comparisons? Is there a need for similar tests with better accuracy? Are surround systems appropriate equipment for measurements in a clinical situation?

- Only 80 of the 10 000 lists it is possible to realize for the five-word sentences are used in “HiST taleaudiometri” at this stage. Because of all the work invested in selecting the words, splitting the sentences into wave-files, adjusting the levels of the individual words and obtaining normal responses – the 400 wave files necessary for generating the next 9920 lists are impatiently waiting on the hard disks, begging to be recycled in new missions!

6.7 What could have been done differently

The following bullet points provide an account of some details that would have been done differently in connection with producing the five-word sentences had it been possible to start afresh with all the experience acquired in the first round:

- Section 2.3.2.1.1 shows the performance-intensity curves for the individual words. Some words have markedly different thresholds or more shallow slopes than others. If we had started out with more than ten words in each category (11, 12 or more) we could have discarded words which require more level adjustment or have an unconstructive influence on the mean value of the slope of the words.

- Section 2.2.2 describes the splitting of the recorded sentences both by the Wagener method and the diphone method. When realizing the sentences according to the diphone methods both of the wave files split by the diphone splitting points are needed, and the splitting points in the Wagener method are needed for level adjustments of the single words. A new routine should be made for decisions involving both the Wagener and diphone splitting points in the same Matlab tool.

- Section 2.2.2.5 describes level adjustments made to the wave files before generating the sentences used for initial measurement of the performance-intensity curves of the individual words. All the wave files were normalized to the same level. This normalization guaranteed that all the generated sentences had the same level. Since the normalization was made on wave files which contained the last part of one word and the initial part of the following word (Table 2.5), the level of each word did not need to be the same. A result of this is that for each word we want to generate, we have ten recordings of the initial part and ten recordings of the last part which can be concatenated in 100 different combinations. The correct combination is determined by the preceding and following word in the sentence to be generated. A better procedure for the normalization prior to the initial measurement of the performance-intensity curve of each word would be: First, normalize all the 100 recorded sentences to the same level. Second, normalize the ten initial parts of each word to the mean value of these ten initial parts. Finally, use the same procedure and normalize the ten last parts of each word to the mean value of these ten last parts. This would guarantee that all the 100 combinations possible for generating each word would have the same level both before the initial measurement of the performance-intensity curve and after the level adjustment necessary to normalize all the words to the same signal-to-noise threshold. The slope of the words can be expected to be steeper when following this procedure, since all of the 100 combinations for generating a word will have the same level.

6.8 Conclusion

"HiST taleaudiometri" is a new speech audiometry set in Norwegian which has recently been made available in Norway. Extensive testing has been performed on young normal-hearing persons in order to adjust the material for the intended purpose. However, only the future can tell how well the material will be received and how well it will function in the audiological clinical institutions in Norway.

Appendix A

Score protocol for first field test

Table A.1 Score protocol for first field test.

dato		testpersonnr.			øre
sted		alder			
rom		kjønn			Nivå:
utstyr		dialektbakgrunn			
operatør		Normalthørende?			
		(Hvis ikke normalthørende angi terskler på baksiden for testøret)			
Før start må kalibreringsnivået sjekkes uetn testperson, spill spor 85 juster til 0 på audiometeret.					
Sett strek under alle ord som blir oppfattet korrekt					
CD:	**A**	**Antall riktige**			**Antall riktige**
1	Hedda tok fem lette duker.		**Nå er du halvferdig. Det kan være lurt**		
2	Ingvild vant elleve nye vanter		**og ta en kort pause, spørre**		
			testpersonen om det går bra og		
Ta en pause her og hør etter om			**si fra at vi er halvferdig.**		
styrken er passe sterk?			42	Hedda grep åtte mørke vanter.	
Hvis ikke juster og prøv fra starten igjen.			43	Eivind ser to store duker.	
3	Jonas grep tolv lyse kasser.		44	Benjamin eide atten gamle duker.	
4	Thea låner tre fine kurver.		45	Thea har to store penner.	
5	Hedda eide elleve nye vanter.		46	Thea vant fem vakre kurver.	
6	Malin tok seks store boller.		47	Magnus tok tre lette knapper.	
7	Thomas viser fem hele skåler.		48	Jonas viser fire mørke kasser.	
8	Magnus flytter åtte mørke knapper.		49	Eivind vant åtte fine skåler.	
9	Ida ser fire svarte penner.		50	Magnus tok sju nye vanter.	
10	Ingvild ga sju vakre luer.		51	Ingvild eide fire hele ringer.	
11	Eivind vant atten lette duker.		52	Jonas har atten fine penner.	
12	Benjamin har to gamle ringer.		53	Hedda låner atten lette luer.	
13	Thea eide tre fine knapper.		54	Thea ga tolv vakre knapper.	
14	Malin låner åtte hele skåler.		55	Benjamin eide tolv vakre knapper.	
15	Eivind viser fire store penner.		56	Magnus ser fem nye kasser.	
16	Ida vant tolv mørke luer.		57	Malin ga tolv store vanter.	
17	Benjamin har elleve nye boller.		58	Ingvild låner tre lette boller.	
18	Hedda tok fem lette duker.		59	Jonas tok sju mørke boller.	
19	Jonas ser atten vakre vanter.		60	Benjamin viser sju svarte boller.	
20	Magnus grep to gamle ringer.		61	Thomas grep seks hele ringer.	
21	Ingvild ga sju lyse kurver.		62	Malin flytter fire gamle kurver.	
22	Thomas flytter seks svarte kasser.		63	Thomas ga tolv gamle skåler.	
23	Magnus ga seks lette duker.		64	Eivind ser åtte fine knapper.	
24	Thea tok elleve mørke vanter.		65	Magnus viser seks lette kurver.	
25	Ingvild har fem store knapper.		66	Hedda vant to lyse ringer.	
26	Eivind grep åtte gamle ringer.		67	Malin ga atten svarte duker.	
27	Malin vant tre vakre luer.		68	Eivind viser to store duker.	
28	Thomas låner atten nye kasser.		69	Ida har elleve fine vanter.	
29	Ida ser to svarte kurver.		70	Ida låner seks lyse luer.	
30	Jonas flytter sju lyse penner.		71	Ida grep fem lyse kurver.	
31	Benjamin eide tolv fine skåler.		72	Benjamin vant sju svarte skåler.	
32	Hedda viser fire hele boller.		73	Hedda eide elleve nye vanter.	
33	Thea har elleve vakre kasser.		74	Benjamin har to gamle ringer.	
34	Thomas ser elleve gamle ringer.		75	Magnus flytter åtte mørke knapper.	
35	Ingvild eide tre lyse penner.		76	Eivind vant atten lette duker.	
36	Jonas tok åtte mørke boller.		77	Thomas viser fem hele skåler.	
37	Malin flytter elleve nye kasser.		78	Ingvild ga sju vakre luer.	
38	Hedda flytter seks hele luer.		79	Malin tok seks store boller.	
39	Ingvild låner fire hele luer.		80	Thea låner tre fine kurver.	
40	Ida flytter fem svarte skåler.		81	Jonas grep tolv lyse kasser.	
41	Thomas grep tre nye penner.		82	Ida ser fire svarte penner.	
(Hvis du lurer på noe ta kontakt: jon.oygarden@hist.no, tel:73559176, hjem:73511761, mob:92613883)					

Appendix B

Results from first field test

Table B.1 shows the score results in per cent for each word from the first field test.

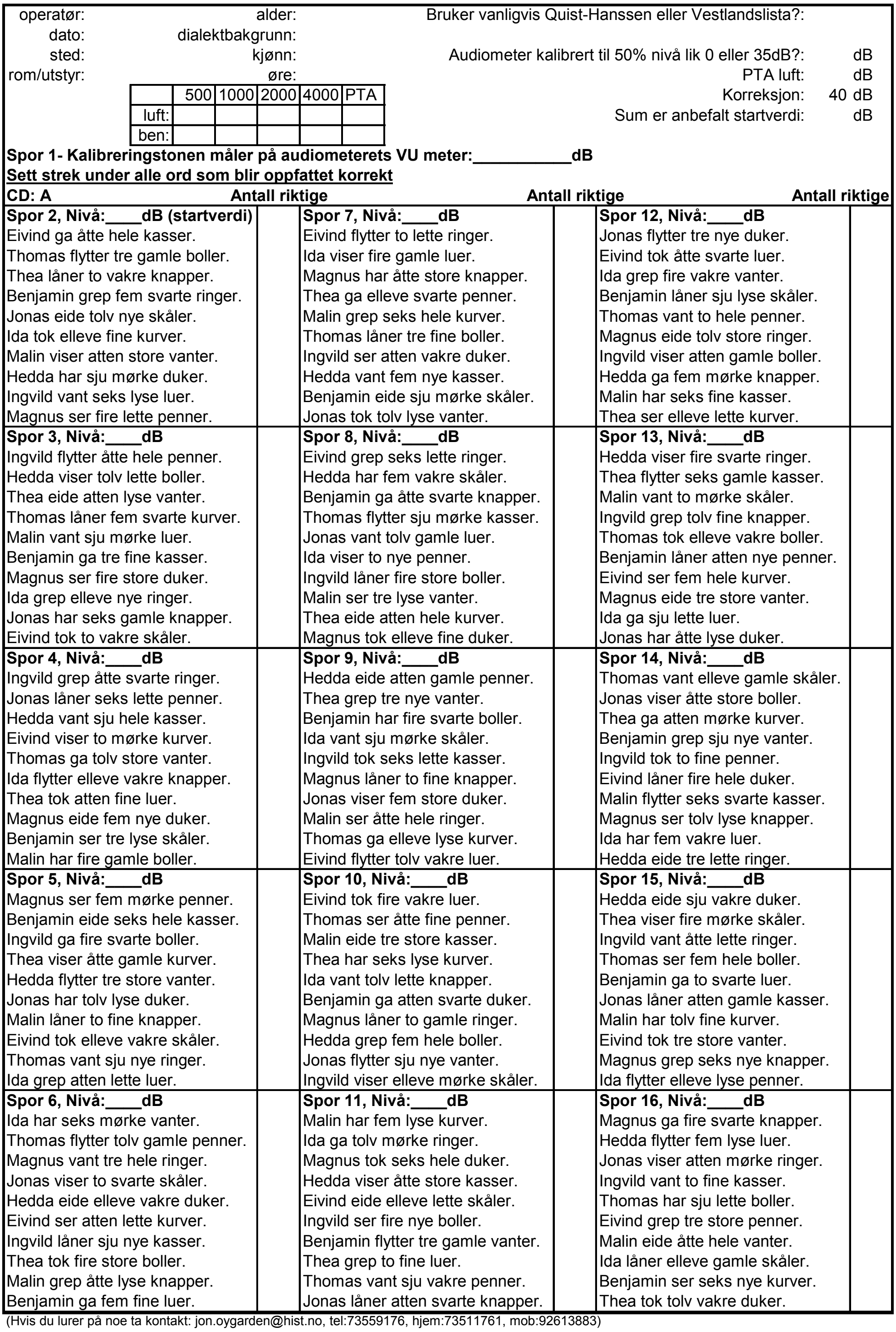
Table B.1 The score results in per cent for each word from the first field test.

# measured	30	18	33	26	30	18	33	26	30	18	33	26	30	18	33	26
Protocol set	A!	B!	C!	D!	A!	B!	C!	D!	A!	B!	C!	D!	A!	B!	C!	D!
SNR [dB]	**-12**	**-11**	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Hedda	0	0	18	4	87	67	52	38	37	89	64	100	97	100	91	96
Ida	0	0	15	8	23	39	76	8	53	100	100	65	90	100	100	100
Malin	0	0	33	31	7	44	79	88	50	94	73	92	97	100	97	100
Ingvild	0	0	6	4	20	28	42	4	43	11	48	85	57	28	82	77
Thea	0	0	24	8	10	22	73	85	80	78	82	81	97	100	100	100
Benjamin	0	0	6	15	7	11	85	62	57	72	82	96	100	100	100	100
Jonas	23	6	6	4	17	39	24	46	87	94	91	100	93	94	100	100
Thomas	3	33	3	58	60	67	88	65	73	100	97	88	100	100	100	92
Magnus	0	0	6	81	90	56	82	92	80	94	97	96	100	100	100	100
Eivind	0	0	3	4	3	28	15	31	33	78	76	54	80	83	88	92
ga	3	0	0	0	0	0	0	46	50	39	76	96	33	100	94	77
grep	0	0	0	0	7	28	27	38	30	50	55	69	50	89	94	69
ser	3	0	0	0	27	50	42	23	87	28	39	96	100	100	94	100
vant	0	0	21	0	50	0	36	8	67	50	82	81	97	89	94	100
låner	0	0	0	0	0	11	30	4	60	17	24	88	100	89	97	96
eide	0	0	0	38	7	11	39	19	30	50	82	96	100	56	100	92
flytter	0	0	6	0	13	0	3	0	27	33	73	65	73	83	82	100
viser	0	0	3	4	33	28	64	62	60	89	73	100	87	100	100	100
har	7	0	27	42	17	33	33	58	80	78	85	88	100	83	97	100
tok	0	0	3	15	7	0	27	23	10	78	79	65	100	89	97	100
to	0	0	15	4	7	11	3	23	80	44	76	81	93	67	85	100
tre	0	0	6	8	20	0	79	69	77	89	85	88	97	94	97	100
fire	3	0	18	50	0	44	48	27	83	89	82	96	100	94	91	96
fem	0	22	3	19	17	50	55	65	87	67	70	92	100	89	97	100
seks	70	28	67	73	60	89	97	88	93	100	94	92	97	100	91	100
sju	17	0	6	0	70	72	76	85	73	78	100	96	100	94	100	100
åtte	0	0	0	8	13	11	30	27	50	50	73	88	100	100	97	96
elleve	0	0	6	4	7	6	24	73	77	61	85	85	97	72	100	92
tolv	3	0	0	4	10	28	64	8	57	89	100	92	87	89	100	100
atten	0	0	15	35	3	0	73	27	27	89	55	81	97	100	97	92
gamle	0	0	9	12	27	72	21	85	77	33	100	88	100	100	100	100
hele	0	0	3	0	3	17	36	27	63	61	48	81	93	61	79	85
store	0	0	9	8	7	56	79	73	100	89	91	100	93	94	100	100
nye	3	0	6	8	10	11	61	81	90	67	94	73	100	83	100	100
vakre	0	0	3	27	33	56	73	50	87	56	100	69	83	50	100	96
mørke	3	0	9	0	10	11	12	23	73	56	18	77	100	83	82	100
lyse	0	0	3	15	13	11	76	69	87	89	79	73	100	94	100	100
fine	0	0	9	15	23	6	27	96	70	61	88	92	93	100	97	96
lette	0	6	0	0	7	6	12	31	17	28	79	81	63	83	91	92
svarte	10	0	15	65	53	78	67	96	80	89	91	92	100	100	97	88
knapper	3	0	42	23	23	50	82	88	97	94	85	81	87	100	100	85
boller	13	0	15	4	60	17	61	69	57	100	97	92	100	94	100	100
vanter	13	0	3	38	30	39	82	88	97	94	91	96	100	100	97	100
penner	0	0	3	19	3	0	85	85	93	100	79	96	100	89	88	96
kurver	0	0	6	4	27	6	15	23	33	56	64	81	93	100	97	100
skåler	3	0	18	46	70	28	94	100	93	100	100	96	100	100	94	96
luer	0	0	6	8	0	6	48	35	20	44	67	81	93	89	88	96
duker	0	0	3	0	3	22	12	65	60	72	82	88	97	100	100	96
ringer	0	0	12	31	60	56	58	46	73	72	79	77	90	100	100	92
kasser	27	39	21	77	67	78	64	100	93	78	94	96	100	94	97	100

Appendix C

Score protocols for second field test

Table C.1 Score protocol for second filed test, page 1.



operatør: alder: Bruker vanligvis Quist-Hanssen eller Vestlandslista?:
dato: dialektbakgrunn:
sted: kjønn: Audiometer kalibrert til 50% nivå lik 0 eller 35dB?: dB
rom/utstyr: øre: PTA luft: dB
Korreksjon: 40 dB
Sum er anbefalt startverdi: dB

	500	1000	2000	4000	PTA
luft:					
ben:					

Spor 1- Kalibreringstonen måler på audiometerets VU meter:___________dB

Sett strek under alle ord som blir oppfattet korrekt

CD: A	**Antall riktige**		**Antall riktige**		**Antall riktige**
Spor 2, Nivå:____dB (startverdi)		**Spor 7, Nivå:____dB**		**Spor 12, Nivå:____dB**	
Eivind ga åtte hele kasser.		Eivind flytter to lette ringer.		Jonas flytter tre nye duker.	
Thomas flytter tre gamle boller.		Ida viser fire gamle luer.		Eivind tok åtte svarte luer.	
Thea låner to vakre knapper.		Magnus har åtte store knapper.		Ida grep fire vakre vanter.	
Benjamin grep fem svarte ringer.		Thea ga elleve svarte penner.		Benjamin låner sju lyse skåler.	
Jonas eide tolv nye skåler.		Malin grep seks hele kurver.		Thomas vant to hele penner.	
Ida tok elleve fine kurver.		Thomas låner tre fine boller.		Magnus eide tolv store ringer.	
Malin viser atten store vanter.		Ingvild ser atten vakre duker.		Ingvild viser atten gamle boller.	
Hedda har sju mørke duker.		Hedda vant fem nye kasser.		Hedda ga fem mørke knapper.	
Ingvild vant seks lyse luer.		Benjamin eide sju mørke skåler.		Malin har seks fine kasser.	
Magnus ser fire lette penner.		Jonas tok tolv lyse vanter.		Thea ser elleve lette kurver.	
Spor 3, Nivå:____dB		**Spor 8, Nivå:____dB**		**Spor 13, Nivå:____dB**	
Ingvild flytter åtte hele penner.		Eivind grep seks lette ringer.		Hedda viser fire svarte ringer.	
Hedda viser tolv lette boller.		Hedda har fem vakre skåler.		Thea flytter seks gamle kasser.	
Thea eide atten lyse vanter.		Benjamin ga åtte svarte knapper.		Malin vant to mørke skåler.	
Thomas låner fem svarte kurver.		Thomas flytter sju mørke kasser.		Ingvild grep tolv fine knapper.	
Malin vant sju mørke luer.		Jonas vant tolv gamle luer.		Thomas tok elleve vakre boller.	
Benjamin ga tre fine kasser.		Ida viser to nye penner.		Benjamin låner atten nye penner.	
Magnus ser fire store duker.		Ingvild låner fire store boller.		Eivind ser fem hele kurver.	
Ida grep elleve nye ringer.		Malin ser tre lyse vanter.		Magnus eide tre store vanter.	
Jonas har seks gamle knapper.		Thea eide atten hele kurver.		Ida ga sju lette luer.	
Eivind tok to vakre skåler.		Magnus tok elleve fine duker.		Jonas har åtte lyse duker.	
Spor 4, Nivå:____dB		**Spor 9, Nivå:____dB**		**Spor 14, Nivå:____dB**	
Ingvild grep åtte svarte ringer.		Hedda eide atten gamle penner.		Thomas vant elleve gamle skåler.	
Jonas låner seks lette penner.		Thea grep tre nye vanter.		Jonas viser åtte store boller.	
Hedda vant sju hele kasser.		Benjamin har fire svarte boller.		Thea ga atten mørke kurver.	
Eivind viser to mørke kurver.		Ida vant sju mørke skåler.		Benjamin grep sju nye vanter.	
Thomas ga tolv store vanter.		Ingvild tok seks lette kasser.		Ingvild tok to fine penner.	
Ida flytter elleve vakre knapper.		Magnus låner to fine knapper.		Eivind låner fire hele duker.	
Thea tok atten fine luer.		Jonas viser fem store duker.		Malin flytter seks svarte kasser.	
Magnus eide fem nye duker.		Malin ser åtte hele ringer.		Magnus ser tolv lyse knapper.	
Benjamin ser tre lyse skåler.		Thomas ga elleve lyse kurver.		Ida har fem vakre luer.	
Malin har fire gamle boller.		Eivind flytter tolv vakre luer.		Hedda eide tre lette ringer.	
Spor 5, Nivå:____dB		**Spor 10, Nivå:____dB**		**Spor 15, Nivå:____dB**	
Magnus ser fem mørke penner.		Eivind tok fire vakre luer.		Hedda eide sju vakre duker.	
Benjamin eide seks hele kasser.		Thomas ser åtte fine penner.		Thea viser fire mørke skåler.	
Ingvild ga fire svarte boller.		Malin eide tre store kasser.		Ingvild vant åtte lette ringer.	
Thea viser åtte gamle kurver.		Thea har seks lyse kurver.		Thomas ser fem hele boller.	
Hedda flytter tre store vanter.		Ida vant tolv lette knapper.		Benjamin ga to svarte luer.	
Jonas har tolv lyse duker.		Benjamin ga atten svarte duker.		Jonas låner atten gamle kasser.	
Malin låner to fine knapper.		Magnus låner to gamle ringer.		Malin har tolv fine kurver.	
Eivind tok elleve vakre skåler.		Hedda grep fem hele boller.		Eivind tok tre store vanter.	
Thomas vant sju nye ringer.		Jonas flytter sju nye vanter.		Magnus grep seks nye knapper.	
Ida grep atten lette luer.		Ingvild viser elleve mørke skåler.		Ida flytter elleve lyse penner.	
Spor 6, Nivå:____dB		**Spor 11, Nivå:____dB**		**Spor 16, Nivå:____dB**	
Ida har seks mørke vanter.		Malin har fem lyse kurver.		Magnus ga fire svarte knapper.	
Thomas flytter tolv gamle penner.		Ida ga tolv mørke ringer.		Hedda flytter fem lyse luer.	
Magnus vant tre hele ringer.		Magnus tok seks hele duker.		Jonas viser atten mørke ringer.	
Jonas viser to svarte skåler.		Hedda viser åtte store kasser.		Ingvild vant to fine kasser.	
Hedda eide elleve vakre duker.		Eivind eide elleve lette skåler.		Thomas har sju lette boller.	
Eivind ser atten lette kurver.		Ingvild ser fire nye boller.		Eivind grep tre store penner.	
Ingvild låner sju nye kasser.		Benjamin flytter tre gamle vanter.		Malin eide åtte hele vanter.	
Thea tok fire store boller.		Thea grep to fine luer.		Ida låner elleve gamle skåler.	
Malin grep åtte lyse knapper.		Thomas vant sju vakre penner.		Benjamin ser seks nye kurver.	
Benjamin ga fem fine luer.		Jonas låner atten svarte knapper.		Thea tok tolv vakre duker.	

(Hvis du lurer på noe ta kontakt: jon.oygarden@hist.no, tel:73559176, hjem:73511761, mob:92613883)

Table C.2 Score protocol for second filed test, page 2.

Spor 17, Nivå:____dB		Spor 18, Nivå:____dB		Spor 19, Nivå:____dB	
Eivind eide sju nye kurver.		Hedda flytter fem gamle ringer.		Ida ser fem fine boller.	
Hedda ser to mørke skåler.		Benjamin tok to fine luer.		Malin grep tolv lyse luer.	
Magnus vant åtte vakre kasser.		Jonas har tre store duker.		Thea låner to nye kurver.	
Ingvild låner seks lyse vanter.		Ida grep seks vakre knapper.		Ingvild flytter atten svarte duker.	
Jonas ga tre fine knapper.		Malin ga sju lyse vanter.		Hedda viser elleve mørke penner.	
Thomas tok fem lette boller.		Eivind vant åtte nye penner.		Magnus har tre lette knapper.	
Ida grep fire store ringer.		Ingvild ser atten mørke kurver.		Benjamin vant seks gamle skåler.	
Malin har tolv gamle luer.		Thomas viser tolv hele kasser.		Jonas ga åtte hele kasser.	
Benjamin flytter elleve hele duker.		Thea eide fire lette boller.		Eivind tok sju store vanter.	
Thea viser atten svarte penner.		Magnus låner elleve svarte skåler.		Thomas eide fire vakre ringer.	

Spor 20, Nivå:____dB		**Spor 25, Nivå:____dB**		**Spor 30, Nivå:____dB**	
seks fine knapper.		atten nye luer.		to store ringer.	
elleve mørke boller.		seks gamle kurver.		tolv mørke knapper.	
sju gamle duker.		elleve mørke ringer.		elleve gamle vanter.	
tre lette kasser.		fem lette kasser.		sju svarte boller.	
fire nye luer.		sju lyse vanter.		fem nye kurver.	
atten lyse ringer.		fire svarte boller.		fire vakre skåler.	
fem vakre penner.		tre hele knapper.		seks lyse penner.	
to store skåler.		to fine skåler.		tre lette duker.	
åtte svarte kurver.		åtte store duker.		atten fine luer.	
tolv hele vanter.		tolv vakre penner.		åtte hele kasser.	
Spor 21, Nivå:____dB		**Spor 26, Nivå:____dB**		**Spor 31, Nivå:____dB**	
tre hele ringer.		elleve lyse skåler.		atten lyse skåler.	
fire store knapper.		seks vakre boller.		fem lette boller.	
seks lette boller.		åtte svarte kurver.		åtte nye luer.	
atten lyse skåler.		to fine ringer.		elleve hele knapper.	
sju svarte kasser.		tolv nye knapper.		sju vakre kasser.	
tolv nye luer.		fire gamle luer.		fire gamle ringer.	
to gamle duker.		fem mørke vanter.		tolv mørke duker.	
fem mørke vanter.		tre store penner.		to svarte vanter.	
elleve fine kurver.		atten lette duker.		tre store kurver.	
åtte vakre penner.		sju hele kasser.		seks fine penner.	
Spor 22, Nivå:____dB		**Spor 27, Nivå:____dB**		**Spor 32, Nivå:____dB**	
to nye duker.		to fine duker.		atten vakre boller.	
fire fine boller.		åtte gamle penner.		fem hele skåler.	
fem lette skåler.		elleve nye kasser.		to svarte ringer.	
tre mørke vanter.		seks mørke ringer.		fire lette penner.	
atten lyse penner.		sju store knapper.		tre store kurver.	
åtte vakre ringer.		fire svarte boller.		tolv mørke luer.	
tolv hele kurver.		fem hele vanter.		sju gamle knapper.	
sju store kasser.		atten lyse skåler.		elleve lyse vanter.	
elleve svarte luer.		tre vakre luer.		åtte fine duker.	
seks gamle knapper.		tolv lette kurver.		seks nye kasser.	
Spor 23, Nivå:____dB		**Spor 28, Nivå:____dB**		**Spor 33, Nivå:____dB**	
tolv vakre luer.		tre mørke kasser.		sju vakre ringer.	
atten hele kurver.		atten nye kurver.		tre lyse knapper.	
to nye kasser.		tolv hele knapper.		tolv mørke kasser.	
seks gamle penner.		to vakre penner.		fire fine luer.	
fem lette ringer.		fem lette ringer.		åtte lette boller.	
åtte lyse skåler.		elleve gamle luer.		seks nye vanter.	
tre mørke vanter.		sju fine boller.		atten store penner.	
fire store knapper.		seks lyse duker.		elleve svarte skåler.	
sju fine boller.		fire store skåler.		fem gamle duker.	
elleve svarte duker.		åtte svarte vanter.		to hele kurver.	
Spor 24, Nivå:____dB		**Spor 29, Nivå:____dB**		**Spor 34, Nivå:____dB**	
fem fine skåler.		seks gamle boller.		åtte svarte boller.	
åtte lette luer.		tre lette duker.		elleve gamle duker.	
seks hele knapper.		fem nye luer.		fire lette skåler.	
tre gamle kasser.		atten lyse vanter.		tolv lyse luer.	
tolv mørke ringer.		sju hele knapper.		atten mørke ringer.	
elleve nye vanter.		åtte svarte kasser.		sju fine penner.	
sju svarte duker.		to mørke kurver.		seks vakre vanter.	
fire lyse boller.		tolv vakre penner.		fem store kasser.	
to vakre kurver.		fire store ringer.		tre nye knapper.	
atten store penner.		elleve fine skåler.		to hele kurver.	

Setninger i støy skal måles på startnivå+5dB

Spor 35, Nivå___dB	
Jonas grep tolv lyse kasser.	
Thea låner tre fine kurver.	
Hedda eide elleve nye vanter.	
Malin tok seks store boller.	
Thomas viser fem hele skåler.	
Magnus flytter åtte mørke knapper.	
Ida ser fire svarte penner.	
Ingvild ga sju vakre luer.	
Eivind vant atten lette duker.	
Benjamin har to gamle ringer.	

Enstavelsesord skal måles med den attenuatorsettingen som ga nærmest 50% skår for treordslistene spor 20-34

Spor 36, Nivå___dB

TRE		KAMP		NOT	
GRIS		BIL		KIS	
TOG		LAND		SNAU	
FISK		DU		GNI	
KATT		SYND		HVEM	
HUS		KNAPP		ROT	
DØR		KINN		FROST	
SAKS		HAUG		SILD	
BLOMST		LIV		MAST	
HUND		UR		TUR	
FOT		GLAD		MIN	
AND		HØY		RÅD	
IS		SKYLL			
MANN		SÅ			
MAT		TRÅD			
SPEIL		MENN			
FUGL		FEIL			
FROSK		FLINK			
RING		NØD			
VENN		SKINN			
MOR		SVAK			
TING		TYNN			
SKY		MEST			
ØY		NATT			
GÅ		LEM			
SAND		DISK			
TAU		DUK			
OVN		SKJEGG			
BORD		HAVN			
FOT		BIT			
LÅS		KJØL			
NÅL		VÅR			
FAST		STEIK			
DE		NORD			
MITT		GI			
GRO		KLANG			
SKJELL		REV			
JORD		LJÅ			

Appendix D

Score protocols for second laboratory test September 2007

Table D.1 Score protocol for first laboratory test, page 1.

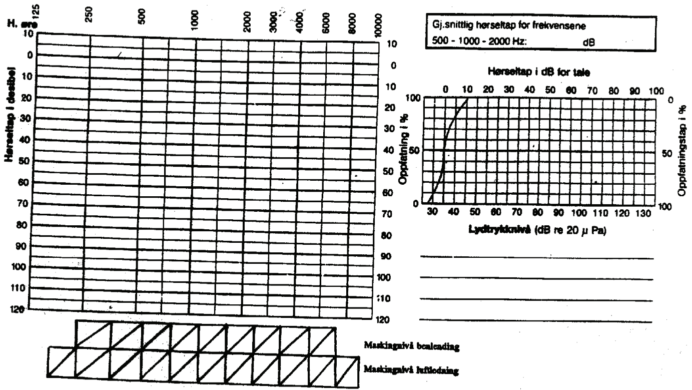
1

cd A

Taleaudiometri PAU2006 tester PAU2007
Sjekk av hvordan den nye HiST taleaudiometri samsvarer med Quist-Hanssen?

alder:	
mann	kvinne
dialektbakgrunn:	

spor 1	Kalibreringstone justeres til 0	utført:
spor 2	Quist-Hanssen tretall uføres på H&V	
spor 3	Quist-Hanssen enstavelsesord uføres på H&V	
spor 4	Quist-Hanssen spondeer uføres på H&V	



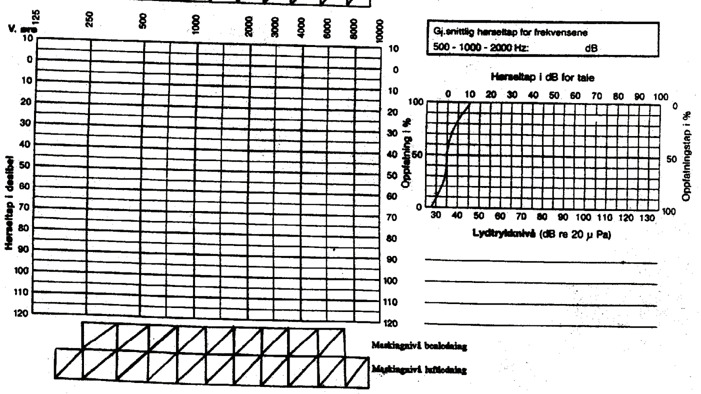
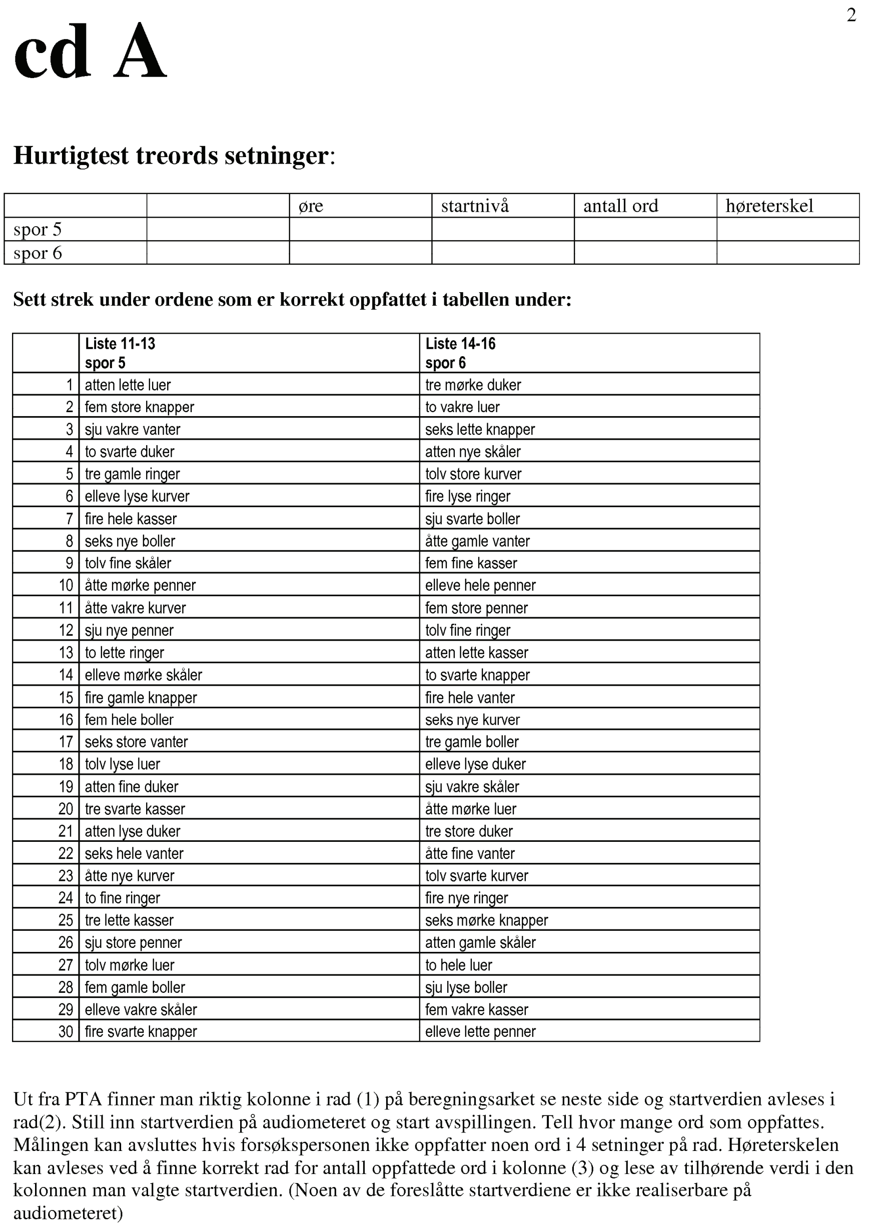


Table D.2 Score protocol for first laboratory test, page 2.



2

cd A

Hurtigtest treords setninger:

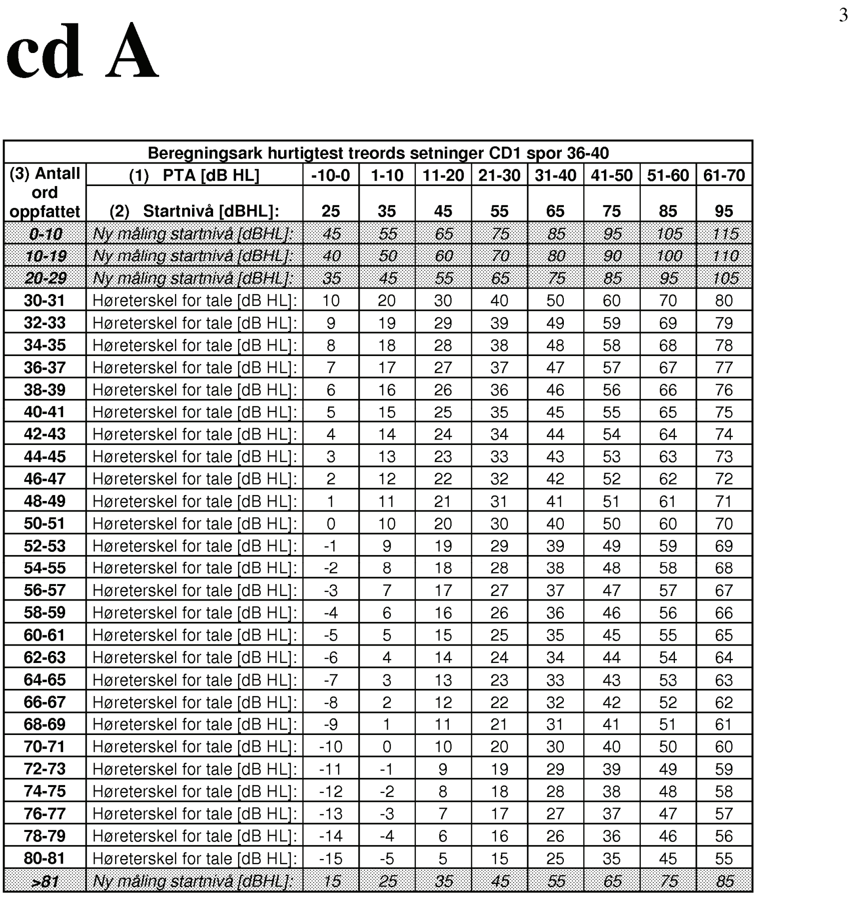
		øre	startnivå	antall ord	høreterskel
spor 5					
spor 6					

Sett strek under ordene som er korrekt oppfattet i tabellen under:

	Liste 11-13 **spor 5**	**Liste 14-16** **spor 6**
1	atten lette luer	tre mørke duker
2	fem store knapper	to vakre luer
3	sju vakre vanter	seks lette knapper
4	to svarte duker	atten nye skåler
5	tre gamle ringer	tolv store kurver
6	elleve lyse kurver	fire lyse ringer
7	fire hele kasser	sju svarte boller
8	seks nye boller	åtte gamle vanter
9	tolv fine skåler	fem fine kasser
10	åtte mørke penner	elleve hele penner
11	åtte vakre kurver	fem store penner
12	sju nye penner	tolv fine ringer
13	to lette ringer	atten lette kasser
14	elleve mørke skåler	to svarte knapper
15	fire gamle knapper	fire hele vanter
16	fem hele boller	seks nye kurver
17	seks store vanter	tre gamle boller
18	tolv lyse luer	elleve lyse duker
19	atten fine duker	sju vakre skåler
20	tre svarte kasser	åtte mørke luer
21	atten lyse duker	tre store duker
22	seks hele vanter	åtte fine vanter
23	åtte nye kurver	tolv svarte kurver
24	to fine ringer	fire nye ringer
25	tre lette kasser	seks mørke knapper
26	sju store penner	atten gamle skåler
27	tolv mørke luer	to hele luer
28	fem gamle boller	sju lyse boller
29	elleve vakre skåler	fem vakre kasser
30	fire svarte knapper	elleve lette penner

Ut fra PTA finner man riktig kolonne i rad (1) på beregningsarket se neste side og startverdien avleses i rad(2). Still inn startverdien på audiometeret og start avspillingen. Tell hvor mange ord som oppfattes. Målingen kan avsluttes hvis forsøkspersonen ikke oppfatter noen ord i 4 setninger på rad. Høreterskelen kan avleses ved å finne korrekt rad for antall oppfattede ord i kolonne (3) og lese av tilhørende verdi i den kolonnen man valgte startverdien. (Noen av de foreslåtte startverdiene er ikke realiserbare på audiometeret)

Table D.3 Score protocol for first laboratory test, page 3.



3

cd A

Beregningsark hurtigtest treords setninger CD1 spor 36-40									
(3) Antall ord oppfattet	**(1) PTA [dB HL]**	**-10-0**	**1-10**	**11-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
	(2) Startnivå [dBHL]:	**25**	**35**	**45**	**55**	**65**	**75**	**85**	**95**
0-10	*Ny måling startnivå [dBHL]:*	*45*	*55*	*65*	*75*	*85*	*95*	*105*	*115*
10-19	*Ny måling startnivå [dBHL]:*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*
20-29	*Ny måling startnivå [dBHL]:*	*35*	*45*	*55*	*65*	*75*	*85*	*95*	*105*
30-31	Høreterskel for tale [dB HL]:	10	20	30	40	50	60	70	80
32-33	Høreterskel for tale [dB HL]:	9	19	29	39	49	59	69	79
34-35	Høreterskel for tale [dB HL]:	8	18	28	38	48	58	68	78
36-37	Høreterskel for tale [dB HL]:	7	17	27	37	47	57	67	77
38-39	Høreterskel for tale [dB HL]:	6	16	26	36	46	56	66	76
40-41	Høreterskel for tale [dB HL]:	5	15	25	35	45	55	65	75
42-43	Høreterskel for tale [dB HL]:	4	14	24	34	44	54	64	74
44-45	Høreterskel for tale [dB HL]:	3	13	23	33	43	53	63	73
46-47	Høreterskel for tale [dB HL]:	2	12	22	32	42	52	62	72
48-49	Høreterskel for tale [dB HL]:	1	11	21	31	41	51	61	71
50-51	Høreterskel for tale [dB HL]:	0	10	20	30	40	50	60	70
52-53	Høreterskel for tale [dB HL]:	-1	9	19	29	39	49	59	69
54-55	Høreterskel for tale [dB HL]:	-2	8	18	28	38	48	58	68
56-57	Høreterskel for tale [dB HL]:	-3	7	17	27	37	47	57	67
58-59	Høreterskel for tale [dB HL]:	-4	6	16	26	36	46	56	66
60-61	Høreterskel for tale [dB HL]:	-5	5	15	25	35	45	55	65
62-63	Høreterskel for tale [dB HL]:	-6	4	14	24	34	44	54	64
64-65	Høreterskel for tale [dB HL]:	-7	3	13	23	33	43	53	63
66-67	Høreterskel for tale [dB HL]:	-8	2	12	22	32	42	52	62
68-69	Høreterskel for tale [dB HL]:	-9	1	11	21	31	41	51	61
70-71	Høreterskel for tale [dB HL]:	-10	0	10	20	30	40	50	60
72-73	Høreterskel for tale [dB HL]:	-11	-1	9	19	29	39	49	59
74-75	Høreterskel for tale [dB HL]:	-12	-2	8	18	28	38	48	58
76-77	Høreterskel for tale [dB HL]:	-13	-3	7	17	27	37	47	57
78-79	Høreterskel for tale [dB HL]:	-14	-4	6	16	26	36	46	56
80-81	Høreterskel for tale [dB HL]:	-15	-5	5	15	25	35	45	55
>81	*Ny måling startnivå [dBHL]:*	*15*	*25*	*35*	*45*	*55*	*65*	*75*	*85*

Table D.4 Score protocol for first laboratory test, page 4.

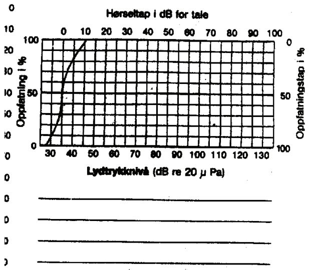
cd A

Treords setninger måling av taleaudiometrikurven:

Før øre og nivå inn i tabellen og sett strek under oppfattede ord.

sp.7, øre:__nivå:__	sp.8, øre:__nivå:__	sp.9, øre:__nivå:__
tre mørke boller	tolv lette knapper	seks gamle skåler
sju svarte skåler	fire store kurver	to lyse boller
tolv store ringer	to gamle vanter	fire lette penner
fem fine penner	tre hele penner	elleve nye ringer
atten nye kasser	åtte lyse ringer	fem svarte kurver
to vakre knapper	fem nye skåler	åtte store duker
fire lyse vanter	atten svarte boller	atten mørke knapper
seks lette kurver	elleve fine kasser	tolv vakre kasser
elleve hele duker	seks vakre luer	tre fine vanter
åtte gamle luer	sju mørke duker	sju hele luer
sp.10, øre:__nivå:__	**sp.11, øre:_nivå:__**	**sp.12, øre:__nivå:__**
fire store skåler	elleve svarte kurver	to gamle ringer
tre hele kurver	atten hele luer	atten svarte duker
atten svarte vanter	sju fine vanter	tolv lette luer
seks vakre duker	fire vakre kasser	fem nye boller
åtte lyse kasser	to store duker	fire store knapper
fem nye luer	fem mørke knapper	seks vakre vanter
elleve fine knapper	åtte lette penner	elleve fine skåler
sju mørke ringer	seks lyse boller	åtte lyse kurver
tolv lette boller	tre nye ringer	tre hele kasser
to gamle penner	tolv gamle skåler	sju mørke penner
sp.13, øre:__nivå:__	**sp.14, øre:_nivå:__**	**sp.15, øre:__nivå:__**
fem svarte luer	atten hele skåler	atten mørke luer
tolv vakre boller	fem mørke kasser	fem svarte knapper
tre fine kurver	sju fine boller	tolv vakre skåler
to lyse penner	tolv gamle kurver	elleve nye kurver
sju hele ringer	tre nye duker	tre fine ringer
åtte store kasser	åtte lette vanter	åtte store penner
atten mørke vanter	to store luer	fire lette kasser
seks gamle duker	elleve svarte penner	to lyse duker
fire lette skåler	seks lyse knapper	sju hele vanter
elleve nye knapper	fire vakre ringer	seks gamle boller
sp.15, øre:__nivå:__		
tre hele luer		
åtte lyse boller		
seks vakre kasser		
elleve fine vanter		
atten svarte kurver		
sju mørke knapper		
fire store duker		
to gamle skåler		
tolv lette penner		
fem nye ringer		

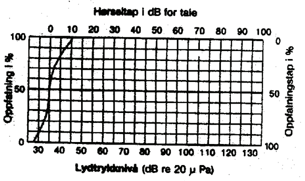
Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB



ingnivå benledning

ingnivå luftledning

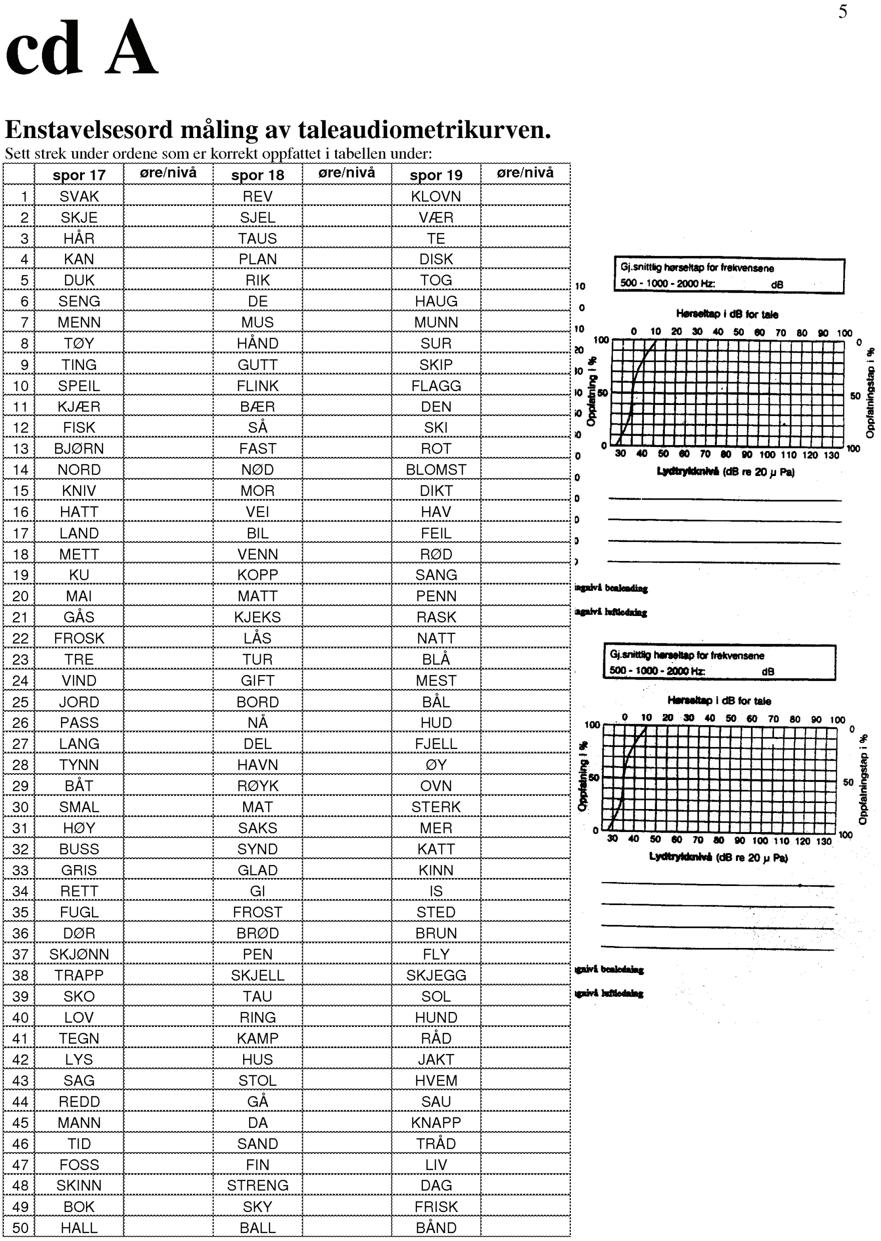
Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB



ingnivå benledning

ingnivå luftledning

Table D.5 Score protocol for first laboratory test, page 5.



5

cd A

Enstavelsesord måling av taleaudiometrikurven.

Sett strek under ordene som er korrekt oppfattet i tabellen under:

	spor 17	øre/nivå	spor 18	øre/nivå	spor 19	øre/nivå
1	SVAK		REV		KLOVN	
2	SKJE		SJEL		VÆR	
3	HÅR		TAUS		TE	
4	KAN		PLAN		DISK	
5	DUK		RIK		TOG	
6	SENG		DE		HAUG	
7	MENN		MUS		MUNN	
8	TØY		HÅND		SUR	
9	TING		GUTT		SKIP	
10	SPEIL		FLINK		FLAGG	
11	KJÆR		BÆR		DEN	
12	FISK		SÅ		SKI	
13	BJØRN		FAST		ROT	
14	NORD		NØD		BLOMST	
15	KNIV		MOR		DIKT	
16	HATT		VEI		HAV	
17	LAND		BIL		FEIL	
18	METT		VENN		RØD	
19	KU		KOPP		SANG	
20	MAI		MATT		PENN	
21	GÅS		KJEKS		RASK	
22	FROSK		LÅS		NATT	
23	TRE		TUR		BLÅ	
24	VIND		GIFT		MEST	
25	JORD		BORD		BÅL	
26	PASS		NÅ		HUD	
27	LANG		DEL		FJELL	
28	TYNN		HAVN		ØY	
29	BÅT		RØYK		OVN	
30	SMAL		MAT		STERK	
31	HØY		SAKS		MER	
32	BUSS		SYND		KATT	
33	GRIS		GLAD		KINN	
34	RETT		GI		IS	
35	FUGL		FROST		STED	
36	DØR		BRØD		BRUN	
37	SKJØNN		PEN		FLY	
38	TRAPP		SKJELL		SKJEGG	
39	SKO		TAU		SOL	
40	LOV		RING		HUND	
41	TEGN		KAMP		RÅD	
42	LYS		HUS		JAKT	
43	SAG		STOL		HVEM	
44	REDD		GÅ		SAU	
45	MANN		DA		KNAPP	
46	TID		SAND		TRÅD	
47	FOSS		FIN		LIV	
48	SKINN		STRENG		DAG	
49	BOK		SKY		FRISK	
50	HALL		BALL		BÅND	

Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB

ingnivå benledning

ngnivå luftledning

Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB

gnivå benledning

gnivå luftledning

Table D.6 Score protocol for first laboratory test, page 6.

cd A

Hurtigtest femords setninger:

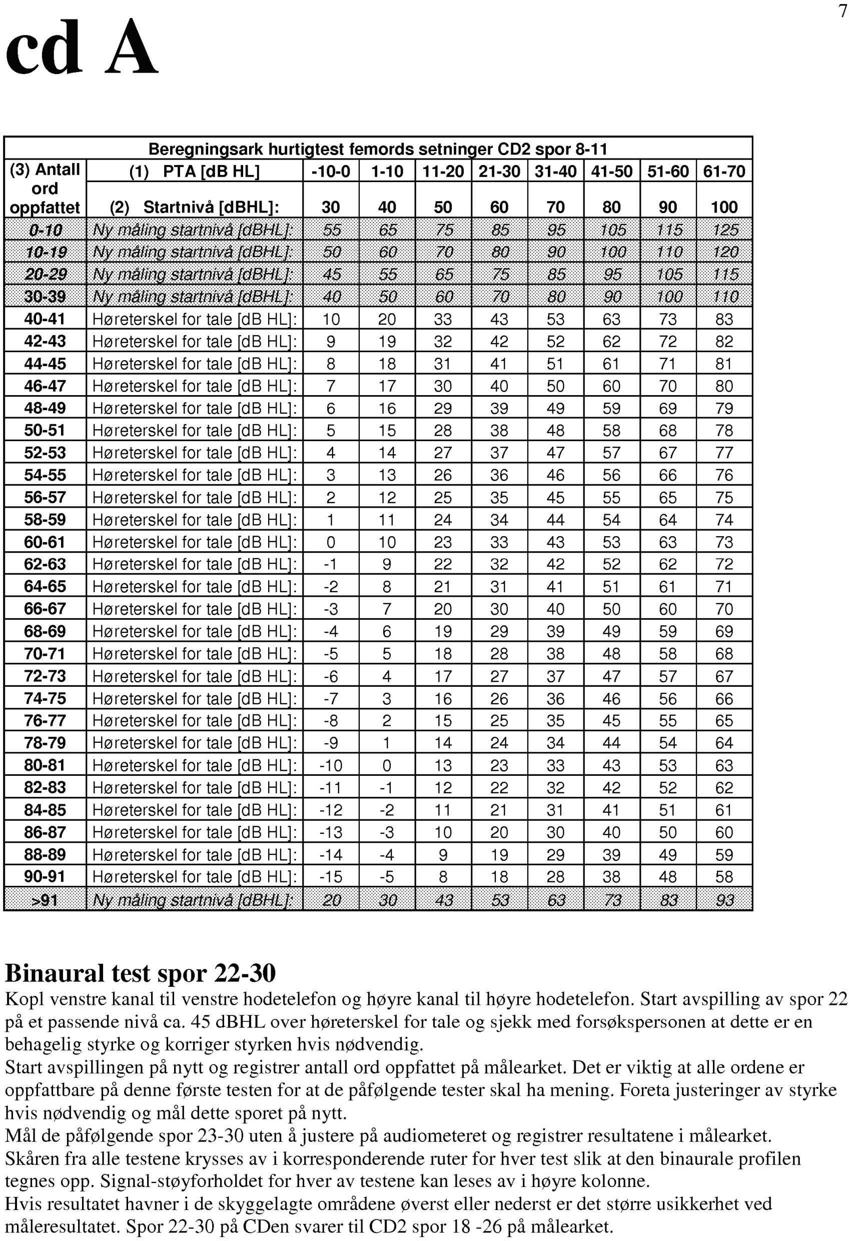
		øre	startnivå	antall ord	høreterskel
spor 20					
spor 21					

Sett strek under ordene som er korrekt oppfattet i tabellen under:

	Liste 85-86 - spor 20	Liste 87-88 - spor 21
1	Malin viser atten lette penner	Thea flytter to vakre duker
2	Hedda eide elleve lyse boller	Ingvild eide atten nye luer
3	Thomas ser seks nye ringer	Magnus ga seks lette boller
4	Ingvild har to svarte kurver	Malin låner tolv store skåler
5	Eivind låner åtte mørke knapper	Jonas har fire lyse kasser
6	Jonas grep fem store duker	Ida vant elleve hele kurver
7	Thea tok tre gamle skåler	Eivind grep sju svarte vanter
8	Ida flytter tolv fine vanter	Thomas tok fem fine knapper
9	Magnus vant sju vakre kasser	Hedda ser åtte gamle penner
10	Benjamin ga fire hele luer	Benjamin viser tre mørke ringer
11	Eivind tok fem lette vanter	Hedda tok atten store knapper
12	Malin ser åtte hele skåler	Thea vant fem lyse kurver
13	Ingvild vant elleve store luer	Eivind har tolv hele kasser
14	Thomas viser tre lyse knapper	Thomas flytter åtte svarte duker
15	Thea låner tolv nye duker	Ida ga to nye boller
16	Hedda ga seks svarte penner	Jonas eide sju lette luer
17	Benjamin eide atten vakre ringer	Benjamin låner seks fine skåler
18	Magnus har fire fine boller	Magnus viser elleve gamle ringer
19	Ida grep sju gamle kurver	Ingvild ser fire mørke penner
20	Jonas flytter to mørke kasser	Malin grep tre vakre vanter

Ut fra PTA finner man riktig kolonne i rad (1) på beregningsarket neste side og startverdien avleses i rad(2). Still inn startverdien på audiometeret og start avspillingen. Tell hvor mange ord som oppfattes. Målingen kan avsluttes hvis forsøkspersonen ikke oppfatter noen ord i 4 setninger på rad. Høreterskelen kan avleses ved å finne korrekt rad for antall oppfattede ord i kolonne (3) og lese av tilhørende verdi i den kolonnen man valgte startverdien. (Noen av de foreslåtte startverdiene er ikke realiserbare på audiometeret)

Table D.7 Score protocol for first laboratory test, page 7.



7

cd A

Beregningsark hurtigtest femords setninger CD2 spor 8-11									
(3) Antall ord oppfattet	**(1) PTA [dB HL]**	**-10-0**	**1-10**	**11-20**	**21-30**	**31-40**	**41-50**	**51-60**	**61-70**
	(2) Startnivå [dBHL]:	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
0-10	*Ny måling startnivå [dBHL]:*	*55*	*65*	*75*	*85*	*95*	*105*	*115*	*125*
10-19	*Ny måling startnivå [dBHL]:*	*50*	*60*	*70*	*80*	*90*	*100*	*110*	*120*
20-29	*Ny måling startnivå [dBHL]:*	*45*	*55*	*65*	*75*	*85*	*95*	*105*	*115*
30-39	*Ny måling startnivå [dBHL]:*	*40*	*50*	*60*	*70*	*80*	*90*	*100*	*110*
40-41	Høreterskel for tale [dB HL]:	10	20	33	43	53	63	73	83
42-43	Høreterskel for tale [dB HL]:	9	19	32	42	52	62	72	82
44-45	Høreterskel for tale [dB HL]:	8	18	31	41	51	61	71	81
46-47	Høreterskel for tale [dB HL]:	7	17	30	40	50	60	70	80
48-49	Høreterskel for tale [dB HL]:	6	16	29	39	49	59	69	79
50-51	Høreterskel for tale [dB HL]:	5	15	28	38	48	58	68	78
52-53	Høreterskel for tale [dB HL]:	4	14	27	37	47	57	67	77
54-55	Høreterskel for tale [dB HL]:	3	13	26	36	46	56	66	76
56-57	Høreterskel for tale [dB HL]:	2	12	25	35	45	55	65	75
58-59	Høreterskel for tale [dB HL]:	1	11	24	34	44	54	64	74
60-61	Høreterskel for tale [dB HL]:	0	10	23	33	43	53	63	73
62-63	Høreterskel for tale [dB HL]:	-1	9	22	32	42	52	62	72
64-65	Høreterskel for tale [dB HL]:	-2	8	21	31	41	51	61	71
66-67	Høreterskel for tale [dB HL]:	-3	7	20	30	40	50	60	70
68-69	Høreterskel for tale [dB HL]:	-4	6	19	29	39	49	59	69
70-71	Høreterskel for tale [dB HL]:	-5	5	18	28	38	48	58	68
72-73	Høreterskel for tale [dB HL]:	-6	4	17	27	37	47	57	67
74-75	Høreterskel for tale [dB HL]:	-7	3	16	26	36	46	56	66
76-77	Høreterskel for tale [dB HL]:	-8	2	15	25	35	45	55	65
78-79	Høreterskel for tale [dB HL]:	-9	1	14	24	34	44	54	64
80-81	Høreterskel for tale [dB HL]:	-10	0	13	23	33	43	53	63
82-83	Høreterskel for tale [dB HL]:	-11	-1	12	22	32	42	52	62
84-85	Høreterskel for tale [dB HL]:	-12	-2	11	21	31	41	51	61
86-87	Høreterskel for tale [dB HL]:	-13	-3	10	20	30	40	50	60
88-89	Høreterskel for tale [dB HL]:	-14	-4	9	19	29	39	49	59
90-91	Høreterskel for tale [dB HL]:	-15	-5	8	18	28	38	48	58
>91	*Ny måling startnivå [dBHL]:*	*20*	*30*	*43*	*53*	*63*	*73*	*83*	*93*

Binaural test spor 22-30

Kopl venstre kanal til venstre hodetelefon og høyre kanal til høyre hodetelefon. Start avspilling av spor 22 på et passende nivå ca. 45 dBHL over høreterskel for tale og sjekk med forsøkspersonen at dette er en behagelig styrke og korriger styrken hvis nødvendig.

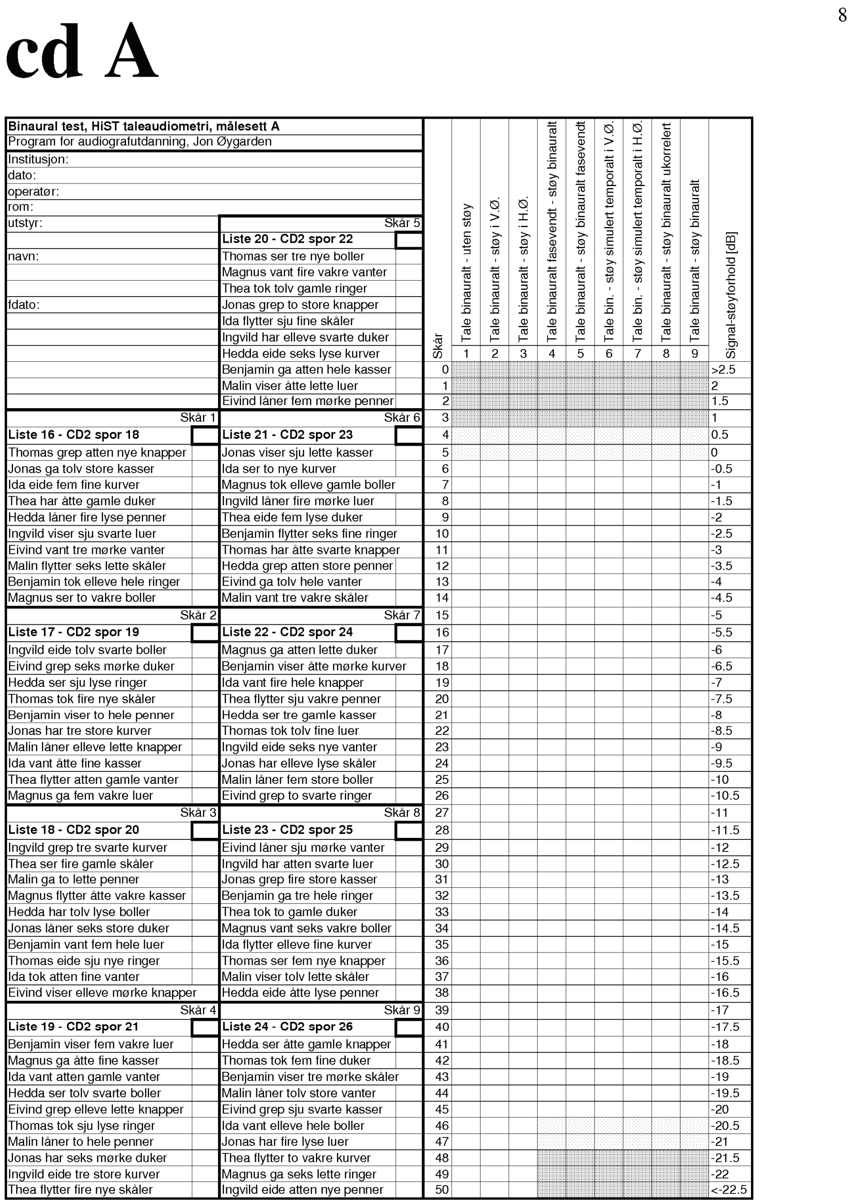
Start avspillingen på nytt og registrer antall ord oppfattet på målearket. Det er viktig at alle ordene er oppfattbare på denne første testen for at de påfølgende tester skal ha mening. Foreta justeringer av styrke hvis nødvendig og mål dette sporet på nytt.

Mål de påfølgende spor 23-30 uten å justere på audiometeret og registrer resultatene i målearket.

Skåren fra alle testene krysses av i korresponderende ruter for hver test slik at den binaurale profilen tegnes opp. Signal-støyforholdet for hver av testene kan leses av i høyre kolonne.

Hvis resultatet havner i de skyggelagte områdene øverst eller nederst er det større usikkerhet ved måleresultatet. Spor 22-30 på CDen svarer til CD2 spor 18 -26 på målearket.

Table D.8 Score protocol for first laboratory test, page 8.



8

cd A

				Skår	Tale binauralt - uten støy	Tale binauralt - støy i V.Ø.	Tale binauralt - støy i H.Ø.	Tale binauralt fasevendt - støy binauralt	Tale binauralt - støy binauralt fasevendt	Tale bin. - støy simulert temporalt i V.Ø.	Tale bin. - støy simulert temporalt i H.Ø.	Tale binauralt - støy binauralt ukorrelert	Tale binauralt - støy binauralt	Signal-støyforhold [dB]
Binaural test, HiST taleaudiometri, målesett A														
Program for audiografutdanning, Jon Øygarden														
Institusjon:														
dato:														
operatør:														
rom:														
utstyr:		Skår 5												
		Liste 20 - CD2 spor 22	□											
navn:		Thomas ser tre nye boller												
		Magnus vant fire vakre vanter												
		Thea tok tolv gamle ringer												
fdato:		Jonas grep to store knapper												
		Ida flytter sju fine skåler												
		Ingvild har elleve svarte duker												
		Hedda eide seks lyse kurver			1	2	3	4	5	6	7	8	9	
		Benjamin ga atten hele kasser		0										>2.5
		Malin viser åtte lette luer		1										2
		Eivind låner fem mørke penner		2										1.5
Skår 1		Skår 6		3										1
Liste 16 - CD2 spor 18	□	**Liste 21 - CD2 spor 23**	□	4										0.5
Thomas grep atten nye knapper		Jonas viser sju lette kasser		5										0
Jonas ga tolv store kasser		Ida ser to nye kurver		6										-0.5
Ida eide fem fine kurver		Magnus tok elleve gamle boller		7										-1
Thea har åtte gamle duker		Ingvild låner fire mørke luer		8										-1.5
Hedda låner fire lyse penner		Thea eide fem lyse duker		9										-2
Ingvild viser sju svarte luer		Benjamin flytter seks fine ringer		10										-2.5
Eivind vant tre mørke vanter		Thomas har åtte svarte knapper		11										-3
Malin flytter seks lette skåler		Hedda grep atten store penner		12										-3.5
Benjamin tok elleve hele ringer		Eivind ga tolv hele vanter		13										-4
Magnus ser to vakre boller		Malin vant tre vakre skåler		14										-4.5
Skår 2		Skår 7		15										-5
Liste 17 - CD2 spor 19	□	**Liste 22 - CD2 spor 24**	□	16										-5.5
Ingvild eide tolv svarte boller		Magnus ga atten lette duker		17										-6
Eivind grep seks mørke duker		Benjamin viser åtte mørke kurver		18										-6.5
Hedda ser sju lyse ringer		Ida vant fire hele knapper		19										-7
Thomas tok fire nye skåler		Thea flytter sju vakre penner		20										-7.5
Benjamin viser to hele penner		Hedda ser tre gamle kasser		21										-8
Jonas har tre store kurver		Thomas tok tolv fine luer		22										-8.5
Malin låner elleve lette knapper		Ingvild eide seks nye vanter		23										-9
Ida vant åtte fine kasser		Jonas har elleve lyse skåler		24										-9.5
Thea flytter atten gamle vanter		Malin låner fem store boller		25										-10
Magnus ga fem vakre luer		Eivind grep to svarte ringer		26										-10.5
Skår 3		Skår 8		27										-11
Liste 18 - CD2 spor 20	□	**Liste 23 - CD2 spor 25**	□	28										-11.5
Ingvild grep tre svarte kurver		Eivind låner sju mørke vanter		29										-12
Thea ser fire gamle skåler		Ingvild har atten svarte luer		30										-12.5
Malin ga to lette penner		Jonas grep fire store kasser		31										-13
Magnus flytter åtte vakre kasser		Benjamin ga tre hele ringer		32										-13.5
Hedda har tolv lyse boller		Thea tok to gamle duker		33										-14
Jonas låner seks store duker		Magnus vant seks vakre boller		34										-14.5
Benjamin vant fem hele luer		Ida flytter elleve fine kurver		35										-15
Thomas eide sju nye ringer		Thomas ser fem nye knapper		36										-15.5
Ida tok atten fine vanter		Malin viser tolv lette skåler		37										-16
Eivind viser elleve mørke knapper		Hedda eide åtte lyse penner		38										-16.5
Skår 4		Skår 9		39										-17
Liste 19 - CD2 spor 21	□	**Liste 24 - CD2 spor 26**	□	40										-17.5
Benjamin viser fem vakre luer		Hedda ser åtte gamle knapper		41										-18
Magnus ga åtte fine kasser		Thomas tok fem fine duker		42										-18.5
Ida vant atten gamle vanter		Benjamin viser tre mørke skåler		43										-19
Hedda ser tolv svarte boller		Malin låner tolv store vanter		44										-19.5
Eivind grep elleve lette knapper		Eivind grep sju svarte kasser		45										-20
Thomas tok sju lyse ringer		Ida vant elleve hele boller		46										-20.5
Malin låner to hele penner		Jonas har fire lyse luer		47										-21
Jonas har seks mørke duker		Thea flytter to vakre kurver		48										-21.5
Ingvild eide tre store kurver		Magnus ga seks lette ringer		49										-22
Thea flytter fire nye skåler		Ingvild eide atten nye penner		50										<-22.5

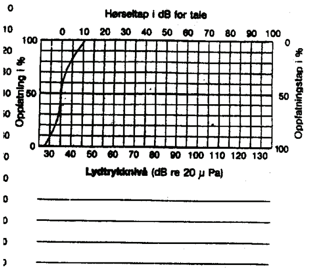
Table D.9 Score protocol for first laboratory test, page 9.

cd A

Ny tretallsliste

spor 31	Quist-Hanssen reviderte tretall uføres på H&V	

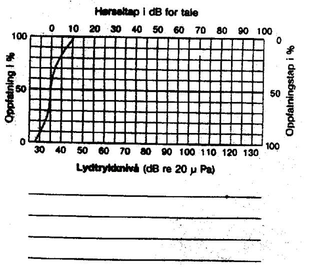
Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB



ngnivå benledning

ngnivå luftledning

Gj.snittlig hørseltap for frekvensene
500 - 1000 - 2000 Hz: dB



gnivå benledning

gnivå luftledning

Appendix E

Monosyllabic words

The following pages show tables of the monosyllabic words.
Table E.1 shows all the words which where evaluated and recorded. The final selections of words are marked with bold typeface in column 2. The words excluded for use are marked with an X in the rightmost column, the reason for exclusion is marked with bold typeface in one or more of the columns: Not included Oxford 3000, Norwegian Google pages 2006-09-20 (low score) and /or per cent recognized (low or high score).
Table E.2-E.4 show nine lists of 50 monosyllabic words included in “HiST taleaudiometri”.

Table E.1 The monosyllabic words evaluated for selection.

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number of tests in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
1	**TRE**	tree		24 000 000	16	69		x		x	
2	**BÆR**		berry	686 000	13	69		x	x	x	
3	**SENG**	bed		887 000	13	54		x			
4	**BALL**	ball		1 490 000	13	69		x			
5	**SKO**	shoe		4 070 000	11	55		x			
6	**GRIS**	pig		432 000	13	54		x	x	x	
7	**TOG**	train		1 870 000	14	21	D	x			
8	**FISK**	fish		4 240 000	16	81	E	x		x	
9	**KOPP**	cup		470 000	13	69		x			
10	**BRØD**	bread		1 520 000	13	46		x			
11	**FLY**	plane		3 700 000	13	54		x	x		
12	**MUS**	mouse		4 480 000	12	50		x			
13	**KATT**	cat		1 240 000	14	93	E	x	x	x	
14	**HUS**	house		13 700 000	15	80	E	x			
15	**DØR**	door		2 750 000	16	50		x			
17	**BOK**	book		7 860 000	13	77		x	x	x	
18	**KLOVN**		clown	163 000	13	46		x			
19	**HEST**	horse		1 900 000	13	85	E	x	x	x	
20	**KU**	cow		711 000	12	50		x			
21	**SAKS**	scissors		602 000	14	57		x			
22	**BLOMST**	flower		627 000	15	60		x			
23	**HUND**	dog		2 440 000	16	69		x	x	x	
24	**TRAPP**	stair		333 000	13	54		x			
25	**SAU**	sheep		607 000	13	69		x		x	
26	**BJØRN**	bear		9 080 000	13	77		x			
27	**KJEKS**	cookie		260 000	12	75		x			
29	**FOT**	foot		1 180 000	14	43		x	x	x	
30	**AND**		duck	10 800 000	15	47		x			
31	**IS**	ice		7 140 000	16	63		x			
32	**GUTT**	boy		3 030 000	13	77		x	x	x	
34	**LAM**		lamb	833 000	13	38	D	x			

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
35	**STOL**	chair		889 000	13	69		x			
36	**KNIV**	knife		481 000	12	25	D	x			
37	**MANN**	man		9 110 000	14	36	D	x	x	x	
38	**MAT**	food		16 000 000	15	60		x			
39	**SPEIL**	mirror		736 000	16	75		x			
40	**BUSS**	bus		2 680 000	13	85	E	x			
41	**LYS**	light		6 760 000	13	85	E	x			
42	**HÅND**	hand		3 640 000	13	46		x			
43	**BÅT**	boat		11 000 000	12	58		x	x		
44	**FUGL**	bird		769 000	14	21	D	x			
47	**FROSK**		frog	211 000	15	80	E	x			
48	**RING**	ring		6 360 000	16	63		x		x	
49	**MUNN**	mouth		1 040 000	13	46		x		x	
50	**FLAGG**	flag		3 370 000	13	46		x			
52	**DAG**	day		39 100 000	13	62			x	x	
53	**HATT**	hat		12 400 000	12	50			x	x	
54	**VENN**	friend		30 600 000	14	71			x	x	
55	**MOR**	mother		3 830 000	15	27	D		x	x	
57	**TING**	thing		11 100 000	16	69			x	x	
59	**BÅL**		campfire /bonfire	305 000	18	22	D	x	x	x	
61	**RØD**	red		4 850 000	13	54			x	x	
66	**VEI**	road		13 100 000	12	67			x	x	
68	**SKY**	cloud		1 430 000	14	50			x	x	
69	**ØY**	island		937 000	15	73			x	x	
71	**PENN**	pen		440 000	13	77			x	x	
72	**SOL**	sun		4 410 000	13	69			x	x	
73	**HAV**	ocean		2 230 000	13	62			x	x	
74	**FIN**	fine		5 590 000	12	58			x	x	
75	**SAND**	sand		2 200 000	14	50			x	x	
76	**TAU**	rope		649 000	15	47			x	x	
79	**OVN**	oven		321 000	16	75			x	x	
83	**REDD**	afraid		3 200 000	13	69			x	x	

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
84	**SANG**	song		2 980 000	13	62			x	x	
88	**LANG**	long		10 300 000	13	62			x	x	
90	**FJELL**	mountain		3 200 000	12	67			x		
91	**BORD**	table		2 990 000	14	36	D		x	x	
93	**LÅS**	lock		550 000	16	88	E		x	x	
95	**SKJE**	spoon		6 030 000	13	100	E		x	x	
96	**RIK**	rich		1 220 000	13	31	D		x	x	
97	**SUR**	sour		1 200 000	13	62			x	x	
99	**VIND**	wind		2 410 000	12	50			x	x	
101	NÅL	needle		**130 000**	14	**14**				x	x
102	**FAST**	tight		7 890 000	15	60				x	
103	**DE**	they		62 000 000	16	44				x	
105	**RETT**	right/straight		23 900 000	13	69				x	
106	MILD	mild		420 000	13	**31**				x	x
107	**GÅ**	go/walk		38 700 000	13	46			x	x	
111	**SAG**	saw		315 000	12	67				x	
112	MITT	mine		11 200 000	14	**21**				x	x
113	GRO	grow		2 000 000	15	**20**				x	x
114	**SKJELL**	shell		517 000	16	69				x	
115	**POST**	post		32 800 000	13	69				x	
116	**HVIT**	white		4 170 000	13	38				x	
117	**DEN**	it		81 500 000	13	38				x	
118	KJØTT	meat		1 480 000	12	**100**				x	x
119	**JORD**	earth		3 550 000	14	36	D			x	
122	**KAMP**	fight		5 800 000	15	53				x	
123	**BIL**	car		19 300 000	16	69		x	x	x	
124	FLOKK		**flock**	**257 000**	13	62				x	x
125	**STERK**	strong		6 080 000	13	77				x	
126	LODD	ticket		**230 000**	13	46				x	x
127	**BRUN**	brown		2 340 000	12	58				x	
129	**LAND**	country		12 300 000	14	50				x	

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
131	DU	you		71 000 000	15	**13**				x	x
133	**SYND**	pity/sorry		1 530 000	16	81	E			x	
134	KAM		**comb**	714 000	13	54				x	x
135	FRI	free		5 350 000	13	**23**				x	x
137	ÅR	year		44 000 000	13	**23**				x	x
138	DET	it		122 000 000	12	**33**				x	x
140	**KNAPP**	button		2 200 000	14	21	D		x	x	
141	**KINN**	cheek		205 000	15	67				x	
142	**HAUG**	hill/pile	heap	1 600 000	16	56				x	
143	**PEN**	nice		1 700 000	13	62				x	
144	**GIFT**	poison/married		1 420 000	13	69				x	
145	REIN		**reindeer**	629 000	13	69				x	x
146	STI	path		843 000	12	**33**				x	x
148	**LIV**	life		12 600 000	14	43				x	
149	UR	watch	scree	964 000	15	**27**				x	x
150	**GLAD**	glad/happy		3 330 000	16	44				x	
151	**SKI**		ski	6 380 000	13	85	E	x	x	x	
152	**BLÅ**	blue		5 370 000	13	62				x	
153	**STED**	place		26 000 000	13	38	D			x	
155	**HØY**	high/tall/loud		9 120 000	14	43				x	
156	SKYLL		**rinse**	**140 000**	15	60				x	x
158	**SÅ**	then	sow	43 700 000	16	81	E			x	
159	**TID**	time		33 200 000	13	54				x	
162	FLAT	flat		1 460 000	13	**23**				x	x
165	**MER**	more		60 400 000	13	54				x	
170	**VÆR**	weather		14 000 000	12	92	E			x	
172	**TRÅD**	thread		4 380 000	14	50				x	
173	**MENN**	men(man pl)		10 900 000	15	67				x	
175	**FEIL**	mistake/fault/error		14 800 000	16	63				x	
179	**SKIP**	ship		4 020 000	13	77				x	
180	**DEL**	part		23 800 000	13	54				x	
181	**TØY**	clothes		338 000	13	62				x	

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
182	**SMAL**	narrow		716 000	12	50				x	
183	**FLINK**	clever		1 280 000	14	64				x	
186	**NØD**	need/want/lack		730 000	15	47				x	
187	**SKINN**	skin		1 070 000	16	69				x	
188	**LOV**	law		17 900 000	13	62				x	
189	**PLAN**	plan		6 920 000	13	46				x	
190	**STRENG**	string/strict		725 000	13	69				x	
191	**HUD**	skin		1 770 000	12	42				x	
192	**SVAK**	weak		2 160 000	14	50				x	
193	**TYNN**	thin		1 280 000	15	40				x	
197	**MEST**	most		31 900 000	16	69				x	
198	**JAKT**	chase/hunting		4 790 000	13	85	E			x	
199	**FRISK**	well/healthy		2 600 000	13	62				x	
200	**HÅR**	hair		2 200 000	13	54				x	
202	**NATT**	night		4 760 000	19	79	E		x	x	
203	LEM		**limb/trapdoor**	272 000	15	40				x	x
204	**DISK**	counter		1 540 000	16	44				x	
205	**NÅ**	now		46 000 000	13	38				x	
206	FET	fat		1 290 000	13	**15**				x	x
207	**HALL**	hall		1 700 000	13	38				x	
208	**DUK**	cloth	tablecloth	527 000	19	37	D		x	x	
210	**SKJEGG**	beard		173 000	15	67				x	
211	**HAVN**	port		2 760 000	16	69				x	
213	**SJEL**	soul		1 100 000	13	85	E			x	
214	DIN	your		41 600 000	13	**23**				x	x
215	**METT**	satisfied		190 000	13	54			x	x	
216	**KAN**	can		76 500 000	12	42				x	
217	BIT	bite		1 210 000	14	**14**				x	x
218	KJØL		**keel**	359 000	15	73				x	x
220	VÅR	spring/our		27 600 000	16	**25**				x	x
221	**PASS**	passport/care		2 910 000	13	77				x	

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
222	NED	down		24 400 000	13	**23**				x	x
223	**GÅS**		goose	275 000	13	46			x	x	
225	FYR	guy/fire		2 260 000	12	**33**				x	x
226	STEIK	joint	roast	**25 300**	14	**21**				x	x
227	**NORD**	north		18 700 000	15	47				x	
228	**GI**	give		20 100 000	16	44				x	
230	**TE**	tea		2 790 000	13	38				x	
231	**RØYK**	smoke		1 140 000	13	54			x	x	
232	**DA**	then		34 100 000	13	38	D			x	
234	PÅ	on		109 000 000	12	**17**				x	x
236	KLANG	sound/ring		184 000	14	**21**				x	x
237	**REV**		fox	1 230 000	15	53		x	x	x	
238	LJÅ		**scythe**	**21 700**	16	56				x	x
239	**SKJÆR**	rock		693 000	13	77				x	
240	**FILM**	film		13 300 000	13	46			x	x	
242	**FOSS**		waterfall	1 220 000	13	69			x	x	
243	**TAUS**	silent		276 000	12	50				x	
244	NOT		**seine/ groove**	2 550 000	14	36				x	x
245	KIS	guy	chap	**133 000**	15	**20**				x	x
246	SNAU		**scant**	**131 000**	16	50				x	x
247	**MATT**	weak	matt/matte	982 000	13	46				x	
248	**DIKT**	poem		4 200 000	13	38	D			x	
250	**SKJØNN**	beautiful/judgement		1 310 000	13	54				x	
251	HÅP	hope		2 420 000	12	**33**				x	x
252	GNI	rub		205 000	14	**7**				x	x
253	**HVEM**	who		15 600 000	15	53				x	
254	**ROT**	root/mess		814 000	16	63				x	
255	**BÅND**	band		968 000	13	54				x	
256	GLATT	smooth		769 000	13	**8**				x	x

File number	Word Included words in bold font	Included in Oxford 3000	Not included Oxford 3000	Norwegian Google pages 2006-09-20	Number tested in listening test	Per cent recognized	D = difficult to recognize E = easy recognized	Included in RC1+2	Included in RC3	Included in Q-H	Excluded
257	VI	we		46 500 000	13	**23**				x	x
259	HEGG		**bird cherry**	**85 000**	12	42				x	x
260	**FROST**		frost	411 000	14	36	D			x	
261	SILD		**herring**	416 000	15	**27**				x	x
262	MAST		**mast**	**134 000**	16	69				x	x
263	**TEGN**	sign		4 570 000	13	38				x	
265	**RASK**	quick		3 340 000	13	62				x	
266	**LYD**	sound		12 100 000	13	46				x	
268	DOKK		**dock**	**136 000**	12	42				x	x
269	**TUR**	tour/journey/walk		10 400 000	14	50				x	
270	MIN	mine		18 400 000	15	**20**				x	x
271	**RÅD**	advice		11 800 000	16	56				x	
272	**BAD**	bath/bathroom		4 070 000	13	62		x		x	
273	HVIS	if		23 800 000	13	**23**				x	x
274	**MAI**	May		17 400 000	13	62				x	
275	**KJÆR**	dear		317 000	12	58				x	

Table E.2 Monosyllabic word lists 1-3.

List 1	File number	List 2	File number	List 3	File number
SVAK	140	REV	108	KLOVN	71
SKJE	125	SJEL	121	VÆR	159
HÅR	61	TAUS	144	TE	145
KAN	66	PLAN	103	DISK	23
DUK	24	RIK	109	TOG	149
SENG	120	DE	19	HAUG	50
MENN	88	MUS	94	MUNN	93
TØY	155	HÅND	60	SUR	139
TING	148	GUTT	45	SKIP	124
SPEIL	134	FLINK	33	FLAGG	32
KJÆR	70	BÆR	13	DEN	21
FISK	30	SÅ	142	SKI	122
BJØRN	5	FAST	26	ROT	111
NORD	96	NØD	97	BLOMST	6
KNIV	73	MOR	92	DIKT	22
HATT	49	VEI	156	HAV	51
LAND	77	BIL	4	FEIL	27
METT	91	VENN	157	RØD	112
KU	75	KOPP	74	SANG	118
MAI	84	MATT	87	PENN	102
GÅS	47	KJEKS	69	RASK	105
FROSK	38	LÅS	83	NATT	95
TRE	151	TUR	153	BLÅ	7
VIND	158	GIFT	42	MEST	90
JORD	64	BORD	9	BÅL	14
PASS	100	NÅ	98	HUD	54
LANG	78	DEL	20	FJELL	31
TYNN	154	HAVN	52	ØY	160
BÅT	16	RØYK	113	OVN	99
SMAL	132	MAT	86	STERK	136
HØY	59	SAKS	116	MER	89
BUSS	12	SYND	141	KATT	67
GRIS	44	GLAD	43	KINN	68
RETT	107	GI	41	IS	62
FUGL	40	FROST	39	STED	135
DØR	25	BRØD	11	BRUN	10
SKJØNN	129	PEN	101	FLY	34
TRAPP	150	SKJELL	127	SKJEGG	126
SKO	130	TAU	143	SOL	133
LOV	80	RING	110	HUND	55
TEGN	146	KAMP	65	RÅD	114
LYS	82	HUS	56	JAKT	63
SAG	115	STOL	137	HVEM	57
REDD	106	GÅ	46	SAU	119
MANN	85	DA	17	KNAPP	72
TID	147	SAND	117	TRÅD	152
FOSS	35	FIN	29	LIV	79
SKINN	123	STRENG	138	DAG	18
BOK	8	SKY	131	FRISK	37
HALL	48	BALL	3	BÅND	15

Table E.3 Monosyllabic word lists 4-5.

List 4	File number	List 5	File number	List 6	File number
SKJÆR	128	GLAD	43	SAKS	116
HEST	53	BUSS	12	LYS	82
LYD	81	SKY	131	SKJEGG	126
FILM	28	DEL	20	TØY	155
LAM	76	KNAPP	72	DA	17
HVIT	58	MAI	84	KLOVN	71
FOT	36	FAST	26	MER	89
BAD	2	HÅR	61	HATT	49
POST	104	ROT	111	BRUN	10
AND	1	VIND	158	FIN	29
SUR	139	SMAL	132	DISK	23
SKJE	125	KATT	67	LÅS	83
RØYK	113	IS	62	HVEM	57
SENG	120	SKJELL	127	TAUS	144
MANN	85	DUK	24	TOG	149
LOV	80	BÅND	15	HALL	48
STOL	137	SAU	119	RØD	112
KAN	66	REV	108	KAMP	65
GUTT	45	TRÅD	152	TRE	151
BJØRN	5	FLAGG	32	SKINN	123
PASS	100	PLAN	103	BAD	2
VÆR	159	HEST	53	FROSK	38
TEGN	146	ØY	160	LIV	79
RING	110	SANG	118	SAND	117
FUGL	40	BORD	9	STED	135
HÅND	60	NØD	97	HØY	59
SKO	130	KOPP	74	LAND	77
BLÅ	7	MUS	94	KU	75
METT	91	HVIT	58	MAT	86
SAG	115	FILM	28	PENN	102
RASK	105	BÆR	13	BRØD	11
SKI	122	JAKT	63	SÅ	142
PEN	101	SKJØNN	129	FLINK	33
TING	148	SPEIL	134	VEI	156
RIK	109	JORD	64	FROST	39
HAVN	52	NÅ	98	SKIP	124
FEIL	27	TE	145	AND	1
DØR	25	REDD	106	SOL	133
BLOMST	6	SVAK	140	RETT	107
GÅS	47	LANG	78	LYD	81
DE	19	GRIS	44	GÅ	46
HUS	56	FISK	30	NATT	95
NORD	96	TYNN	154	FRISK	37
TAU	143	HAUG	50	GI	41
KNIV	73	BÅL	14	MOR	92
MATT	87	MUNN	93	HAV	51
FJELL	31	FOT	36	BALL	3
RÅD	114	DEN	21	HUND	55
STERK	136	TUR	153	POST	104
KINN	68	KJÆR	70	SKJÆR	128

Table E.4 Monosyllabic word lists 7-8.

List 7	File number	List 8	File number	List 9	File number
BOK	8	BÆR	13	OVN	99
SYND	141	SKI	122	FROSK	38
STRENG	138	HUND	55	SKY	131
FLY	34	TYNN	154	TEGN	146
LAM	76	MOR	92	BÅL	14
DAG	18	DISK	23	DØR	25
VENN	157	FJELL	31	PENN	102
MEST	90	GLAD	43	GRIS	44
HUD	54	MUS	94	MATT	87
OVN	99	AND	1	SOL	133
BÅT	16	TAU	143	HATT	49
SJEL	121	SKJE	125	FISK	30
KJEKS	69	LOV	80	SKJÆR	128
TRAPP	150	RASK	105	KU	75
DIKT	22	DA	17	LAM	76
BIL	4	BLOMST	6	BAD	2
MENN	88	VENN	157	MENN	88
FOSS	35	FOT	36	HVIT	58
TID	147	KJÆR	70	TRÅD	152
GIFT	42	DEN	21	SPEIL	134
NORD	96	TØY	155	KINN	68
LÅS	83	SJEL	121	HEST	53
SUR	139	IS	62	REV	108
RØYK	113	FRISK	37	GÅ	46
KNAPP	72	DUK	24	FUGL	40
SAG	115	TRAPP	150	KAMP	65
LANG	78	MUNN	93	NØD	97
HAUG	50	HALL	48	SANG	118
TE	145	STOL	137	ROT	111
HVEM	57	NÅ	98	TAUS	144
BJØRN	5	VEI	156	SKJØNN	129
LYS	82	BUSS	12	SYND	141
SKINN	123	GI	41	LIV	79
STERK	136	STRENG	138	BÅT	16
RIK	109	DIKT	22	JORD	64
GÅS	47	KAN	66	FLINK	33
LAND	77	MER	89	MAT	86
FEIL	27	FLY	34	REDD	106
TUR	153	PASS	100	SAND	117
DE	19	BLÅ	7	SAU	119
PEN	101	HÅND	60	KOPP	74
SÅ	142	HUS	56	VÆR	159
HAVN	52	ØY	160	HØY	59
TING	148	RETT	107	BIL	4
BORD	9	KNIV	73	STED	135
SVAK	140	DAG	18	KLOVN	71
FLAGG	32	RING	110	HÅR	61
MAI	84	MEST	90	METT	91
RØD	112	BALL	3	PLAN	103
POST	104	SKO	130	SAKS	116

Appendix F

Nomenclature for five-word and three-word lists

The nomenclature system for the five-word sentence lists uses the words in Table F.1 as its basis. We have defined a four digit number called a LON-number (List Order Number) in order to identify the 10 000 unique lists it is possible to generate with the material. Each digit gives the amount of cyclic shift for the words in one column. If we start with the names as given in the Name column without any shift, the first digit tells us how many places to shift the verbs, the second digit how many places to shift the numerals, the third digit how many places to shift the adjectives and, finally, the fourth digit how many places to shift the nouns. (An example: LON-number 2106 means that the first sentence is *Hedda ser tre gamle luer* and the second sentence *Ida vant fire hele duker* etc.). By selecting a LON-number we have determined what sentences to use, but before producing the list used for testing we need to randomize the order of our 10 selected sentences.

Table F.1 The words used to generate five-word sentences.

Cyclic shift	No shift	LON 1. digit	LON 2. digit	LON 3. digit	LON 4. digit
Number	*Name*	*verb*	*numeral*	*adjective*	*noun*
0	Hedda	ga	to	gamle	knapper
1	Ida	grep	tre	hele	boller
2	Malin	ser	fire	store	vanter
3	Ingvild	vant	fem	nye	penner
4	Thea	låner	seks	vakre	kurver
5	Benjamin	eide	sju	mørke	skåler
6	Jonas	flytter	åtte	lyse	luer
7	Thomas	viser	elleve	fine	duker
8	Magnus	har	tolv	lette	ringer
9	Eivind	tok	atten	svarte	kasser

A Matlab procedure was prepared in order to measure the levels of the wave files of all the sentences in the 10 000 lists it is possible to realize with the material, transferring the results to a spreadsheet. In the spreadsheet the difference between the sentence with the maximum and the sentence with the minimum level in each list was calculated. The histogram of these differences is presented in Figure F.1

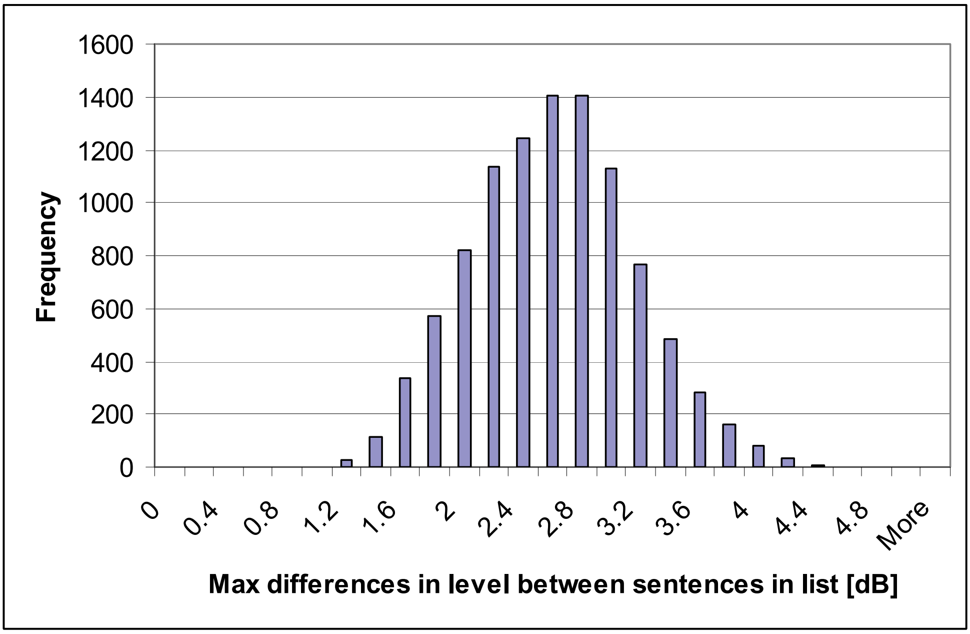


Figure F.1 Histogram of differences between five-word sentences, giving maximum and minimum level in 10 000 lists.

Since we have 10 000 different lists to choose from and only a few lists are needed to make the speech audiometry material, we decided to select first among the lists where the differences in levels between sentences were small. 250 sentences were selected, all of which had differences between maximum and minimum sentence level of less than 1.7 dB. Some lists were excluded in order to avoid too many repetitions of identical word pairs. The 250 selected lists were randomized, and are presented in Table F.2 with the corresponding LON-number. Every realized list used in tests involving five-word sentences in “HiST taleaudiometri” was pulled from this table.

Table F.2 List number for five-word sentences selected, and corresponding LON-number used to generate list.

Lnum	LON	Lnum	LON	Lnum	LON	Lnum	LON	Lnum	LON
1	1491	51	7867	101	6920	151	9269	201	5766
2	6269	52	0999	102	1595	152	4867	202	4464
3	2591	53	1893	103	5963	153	9500	203	2806
4	5906	54	0763	104	2997	154	2381	204	5660
5	1699	55	6464	105	9000	155	9007	205	4821
6	2995	56	4261	106	4491	156	7461	206	9481
7	5266	57	1903	107	2364	157	2695	207	2606
8	9283	58	4964	108	5060	158	1594	208	9981
9	2400	59	2484	109	2497	159	2697	209	4766
10	1481	60	9871	110	5509	160	5903	210	9467
11	0969	61	5966	111	6963	161	1093	211	2871
12	2595	62	0194	112	2369	162	2907	212	9603
13	3467	63	9060	113	1195	163	5403	213	3464
14	2403	64	1606	114	8464	164	3069	214	2500
15	0981	65	2393	115	1694	165	8873	215	3569
16	4263	66	4563	116	4769	166	2060	216	9609
17	2568	67	2605	117	2491	167	9598	217	2091
18	8861	68	7561	118	6967	168	2440	218	4063
19	2891	69	2874	119	1484	169	4266	219	2407
20	5464	70	4568	120	2873	170	9367	220	9900
21	1923	71	2321	121	9699	171	2800	221	4564
22	2109	72	1997	122	4763	172	8866	222	2205
23	5663	73	2900	123	2909	173	1981	223	0873
24	2600	74	6093	124	5568	174	2107	224	4595
25	0490	75	5923	125	1206	175	0983	225	7467
26	8863	76	2870	126	9568	176	2501	226	4260
27	2987	77	0997	127	8821	177	3066	227	5291
28	5920	78	4660	128	3923	178	2609	228	2984
29	4560	79	2805	129	9280	179	0990	229	9097
30	3423	80	4920	130	4873	180	2095	230	0989
31	2106	81	5269	131	5861	181	9873	231	6923
32	4060	82	2100	132	1407	182	5609	232	1994
33	0497	83	9464	133	8961	183	3563	233	9595
34	2981	84	1400	134	2506	184	2101	234	8867
35	6966	85	5761	135	1692	185	1990	235	0191
36	8869	86	0493	136	9821	186	2598	236	9984
37	2906	87	2603	137	2464	187	5094	237	4760
38	5821	88	9920	138	1920	188	2295	238	2373
39	1609	89	2104	139	9581	189	1494	239	5407
40	5261	90	4269	140	4966	190	5769	240	9281
41	9403	91	9903	141	2094	191	2098	241	6876
42	2391	92	2097	142	1891	192	5506	242	9003
43	0986	93	8766	143	2920	193	2494	243	1873
44	5598	94	9106	144	4264	194	9400	244	9861
45	1403	95	5595	145	0195	195	8966	245	2694
46	4497	96	2692	146	2090	196	2699	246	4961
47	0591	97	1497	147	9907	197	6990	247	9407
48	5484	98	2306	148	0996	198	9484	248	8568
49	2505	99	5767	149	1509	199	6997	249	9692
50	9987	100	2903	150	2361	200	0093	250	5763

A similar procedure was followed for the three-word utterances. Here the LON number needed only two digits, one for the adjective and one for the noun, in that order. Figure F.2 shows that the differences between maximum and minimum levels for the utterances within the lists are larger than for the five-word sentences. 80 lists with maximum differences lower than 4.0 dB were selected and randomized. Table F.3 presents this selection. The lists realized in the "HiST taleaudiometri" set were pulled from this table.

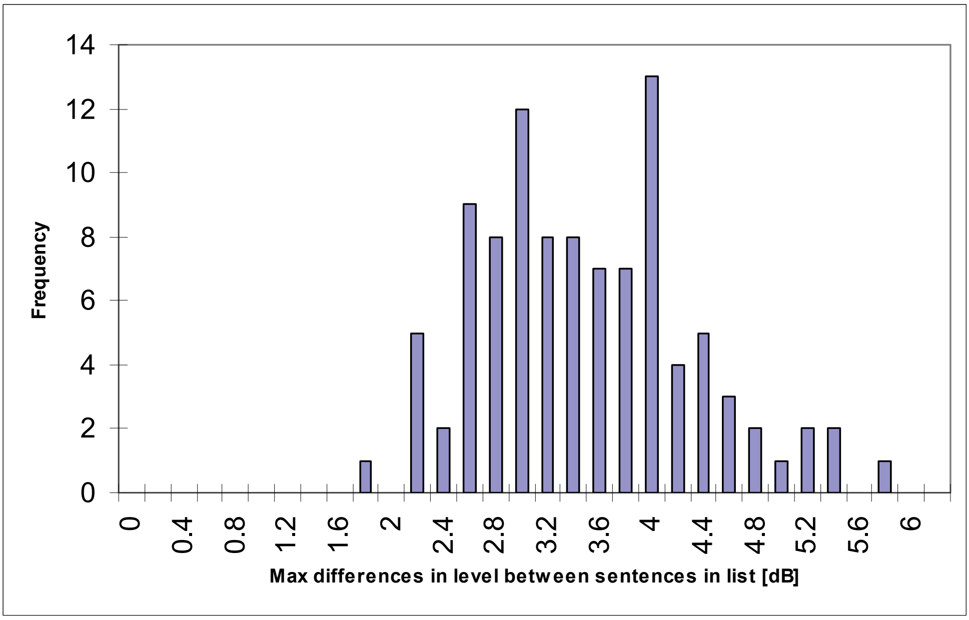


Figure F.2 Histogram of differences between three-word utterances, giving maximum and minimum level in 100 lists.

Table F.3 List number for three-word utterances selected, and corresponding LON-number used to generate list.

Lnum	LON	Lnum	LON	Lnum	LON	Lnum	LON
1	40	21	07	41	73	61	42
2	02	22	71	42	00	62	87
3	61	23	99	43	60	63	43
4	03	24	51	44	33	64	12
5	27	25	79	45	04	65	74
6	08	26	37	46	19	66	13
7	63	27	25	47	21	67	32
8	26	28	18	48	49	68	69
9	67	29	62	49	10	69	23
10	05	30	15	50	39	70	34
11	97	31	20	51	91	71	81
12	88	32	96	52	14	72	22
13	78	33	75	53	95	73	84
14	46	34	35	54	24	74	29
15	90	35	94	55	70	75	64
16	16	36	86	56	17	76	36
17	38	37	92	57	01	77	31
18	93	38	54	58	57	78	98
19	28	39	09	59	72	79	83
20	30	40	76	60	11	80	06

References

Beattie, R. C., 1989. Word Recognition Functions for the CID W-22 Test in Multitalker Noise for Normally Hearing and Hearing-Impaired Subjects. *Journal of Speech and Hearing Disorders,* 54, 20-32.

Beattie, R. C., and Raffin, M. J. M., 1985. Reliability of Threshold, Slope, and PB Max for Monosyllabic Words. *Journal of Speech and Hearing Disorders*, 50, 166-178.

Beattie, R. C., and Warren, V., 1983. Slope Characteristics of CID W-22 Word Functions in Elderly Hearing-Impaired Listeners. *Journal of Speech and Hearing Disorders,* 48, 119-127.

Blume, S. S., and Reeger, B., 1998. Audiometer. *In:* R. Bud and D. J. Warner, eds. *Instruments of Science: An Historical Encyclopedia*. New York: Garland, 39-40.

Boothroyd, A., and Nittrouer, S., 1988. Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, 84(1), 101-114.

Bosman, A., 1992. Review of Speech Audiometric Tests. *In:* B. Kollmeier, ed. *Moderne Verfahren der Sprachaudiometrie.* Heidelberg: Median-Verlag von Killisch-Horn GmbH, 11-34.

Brand, T., and Kollmeier, B., 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6), 2801-2810.

Cardillo G., 2007. *MyFisher24: a very compact routine for Fisher's exact test on 2x4 matrix* [online]. Available from: <http://www.mathworks.com/matlabcentral/fileexchange/19842> [accessed December 12, 2008].

Carhart, R., 1965. Problems in the measurement of speech discrimination. *Archives of Otolaryngology-Head & Neck Surgery*, 82, 253-260.

Carney, E., and Schlauch, R. S., 2007. Critical Difference Table for Word Recognition Testing Derived Using Computer Simulation. *Journal of Speech, Language, and Hearing Research*, 50, 1203–1209.

Cooper, J. C., and Cutts, B. P., 1971. Speech Discrimination in Noise. *Journal of Speech and Hearing Research,* 14, 332-337.

Davis, A. B., and Merzbach, U. C., 1975. Early Auditory Studies - Activities in the Psychology Laboratories of America Universities. *Smithsonian Studies in History and Technology*, 31, 1-52.

Feldmann, H., 2004. 200 Jahre Hörprüfungen mit Sprache, 50 Jahre deutsche Sprachaudiometrie - ein Rückblick. *Laryngo-Rhino-Otol*, 83, 735-742.

Fletcher, H., 1929. *Speech and Hearing*. New York: Van Nostrand.

Fletcher, H., and Steinberg, J. C., 1929. Articulation testing methods. *Bell Sys Tech J.* 8: 806-854.

Gang, R. P., 1976. The Effects of Age on the Diagnostic Utility of the Rollover Phenomenon. *Journal of Speech and Hearing Disorders,* 41, 63-69.

Gelfand, S. A., 2003. Tri-Word Presentations With Phonemic Scoring for Practical High-Reliability Speech Recognition Assessment. *Journal of Speech and Hearing Disorders*, 46, 405-412.

Glasberg, B. R., and Moore, B. C. J., 2002. A model of loudness applicable to time-varying sounds. *J. Audio Eng. Soc.*, 50(5), 331-342.

Gutnick, H. N., and St. John, R., 1982. A model for predicting clinically relevant group differences of open-response tests. *Journal of Speech and Hearing Research*, 25, 468-472.

Hagerman, B., 1976. Reliability in the determination of speech discrimination. *Scandinavian Audiology,* 5, 219-228.

Hagerman, B., 1982. Sentences for testing speech intelligibility in noise. *Scandinavian Audiology,* 11, 79-87.

Hagerman, B., 1989. To find the PB max-Simulations of Levitt's Adaptive Method. Ear and Hearing, 10(5), 323-329.

Hagerman, B., and Kinnefors, C., 1995. Efficient Adaptive Methods for Measuring Speech Reception Thresholds in Quiet and in Noise. *Scandinavian Audiology,* 24, 71-77.

Hastings, A., 2002. *ISO 532 B / DIN 45631 Loudness Matlab code* [online]. Available from: <http://widget.ecn.purdue.edu/~hastinga/Research.htm> [accessed December 13, 2004].

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W., 1952. Development of materials for speech audiometry. *J Speech Hear Disord.* 17, 321-337.

Hornby, A. S., 2005. *Oxford Advanced Learner's Dictionary of Current English,* 7th ed. Oxford: Oxford University Press.

IEC 60318-3, 1998. *Electroacoustics - Simulators of human head and ear - Part 3: Acoustic coupler for the calibration of supra-aural earphones used in audiometry.* Geneva: International Electrotechnical Committee

ISO 532, 1975. *Acoustics – Method for calculating loudness level.* Geneva: International Organization for Standardization.

ISO 8253-3, 1996. *Acoustics – Audiometric test methods – Part 3: Speech audiometry.* Geneva: International Organization for Standardization.

Jerger, J., 1970. Discussion. *In:* C. Røjskjær ed. *Speech Audiometry*, Odense: Second Danavox symposium, 233.

Jusczyk, P. W., and Luce, P. A., 2002. Speech Perception and the Spoken Word Recognition: Past and Present. *Ear & Hearing*, 23(1), 2-40.

Kloster-Jensen, M., 1973. Transients in Testing for Speech Perception. *The Study of Sounds*, 16: 53-67.

Kloster-Jensen, M., 1974. Test Words in Speech Audiometry. *Journal of Audiological Technique*, 13: 158-173.

Kollmeier, B., Brand, T., and Meyer, B., 2008. Perception of Speech and Sound. *In:* J. Benesty, M. M. Sondhi, and Y. Huang, eds. *Springer*

Handbook of Speech Processing. Springer-Verlag Berlin and Heidelberg GmbH & Co, 61-82.

Krause, J. C., and Braida, L. D., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1), 362-378.

Lyregaard, P. E., Robinson, D. W., and Hincliffe, R., 1976. A feasibility study of diagnostic speech audiometry. *NPL Acoustic Report AC73,* National Physical Laboratory, Teddington, Middlesex.

Lyregaard, P., 1997. Towards a theory of speech audiometry tests. *In*: M. Martin ed. *Speech Audiometry,* 2nd ed. London: Whurr, 34-62.

Margolis, R. H., and Millin, J. P., 1971. An Item-Difficulty Based Speech Discrimination Test. *Journal of Speech and Hearing Research,* 14, 865-873.

Martin, F. N., Champlin, C. A., and Perez, D. D., 2000. The Question of Phonetic Balance in Word Recognition Testing. *Journal of the American Academy of Audiology,* 11, 489-493.

Moore, B. C. J., 2003. *An Introduction to the Psychology of Hearing,* 5th ed. San Diego: Academic Press.

Nordgård, T., and Foldvik, A. K., 2001. Reduction of alternative pronunciations in the Norwegian computational lexicon NorKompLeks. *In: Proc. of Eurospeech 01.* Aalborg, DK.

Norsk Språkråd, 2007. [online]. Available from: <http://www.sprakrad.no/English_and_other_languages/English/Norwegian/>, [accessed April 28 2007].

Olsen, W. O., 1990. A Historical Perspective of Hearing Tests of Peripheral Auditory Function. *Journal of the American Academy of Audiology*, 1, 209-216.

Picheny, M. A., Durlach, N. I., and Braida, L. D., 1986. Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing research,* 29, 434-446.

Quist-Hanssen, S., 1965. Hørselmåling. *In:* S. Quist-Hanssen and G. Flottorp, *Hørselskade - undersøkelse og behandling med høreapparat.* 5th ed. Oslo: Aksjeselskapet Erik Høye, 13-24.

Quist-Hanssen, S., 1966. Subjective appraisal and objective assessment of the hearing of speech amongst a group of adults with impaired hearing. *Acta Otolaryngol*., Suppl 224, 177-185.

Quist-Hanssen, S., 1970. Discussion. *In:* C. Røjskjær ed. *Speech Audiometry*, Odense: Second Danavox symposium, 227-228.
Raffin, M. J., and Shafer, D., 1980. Application of a probability model based on the binomial distribution to speech-discriminations scores. *Journal of Speech and Hearing Research,* 23, 570-575.

Schloegl, A., 2004. *Biosig for Octave and Matlab*, [online]. Available from: < http://biosig.sf.net/> [Accessed January 17, 2005].

Slethei, K., 1975. *HS24 en taleaudiometrisk test på fonetisk grunnlag*. Bergen: Report.

Stensby, S., Krokstad, A., and van Dommelen, W., 2002. *Algorithms for hearing aids - Listening tests and applications.* SINTEF report no. STF A02010, Trondheim, Norway.

Sundby, A., 1985. Kalibrering av taleaudiometri. *In: Norsk Audiografforbund/Norsk Teknisk Audiologisk Forening etterutdanningskurs*. 12-15 September 1985, Trondheim.

Thibodeau, L. M., 2007. *Speech audiometry. In*: Roeser, J. R., Valente, M. and Hosford-Dunn, H., ed. *Audiology Diagnosis.* 2nd ed. Thieme New York, 288-313.

Thornton, A. R., and Raffin, M. J. M., 1978. Speech-discrimination scores modelled as a binomial variable. *Journal of Speech and Hearing Research,* 21, 497-506.

Timoney, J., Lysaght, T., Schoenwiesner, M., and MacManus, L., 2004. Implementing loudness models in Matlab. *In: Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX-04)*, October 5-9 2004, Naples, Italy.

Tobias, J. V., 1964. On phonemic analysis of speech discrimination tests. *Journal of Speech and Hearing Research*, 7, 98-100.

University of Bergen, 2003. *Norsk tekstarkiv - De 10000 hyppigste ordformer i norsk - Etter frekvens, ordlistf.zip* [online]. Available from: <http://helmer.aksis.uib.no/nta/> [Accessed January 9, 2003].

University of Bergen, 2007. *Aksis: Norsk tekstarkiv - 10.000 hyppigste ordformer basert på ca 150 millioner ord* [online]. Available from: <http://helmer.aksis.uib.no/nta/> [Accessed March 14, 2007].

Vikør, L. S., 2001. *The Nordic Languages. Their Status and Interrelations*. Oslo: Novus Press 2001.

Vogel, S., 1993. Sensation of Tone, Perception of Sound, and Empiricism - Helmholtz's Physiological Acoustics. *In:* D. Cahan ed. *Hermann Von Helmholtz and the Foundations of Nineteenth-century Science*. University of California Press, Berkeley, 259-290.

Wagener, K., Kühnel, V., and Kollmeier, B., 1999a. Entwiklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie,* 38(1), 4-15.

Wagener, K., Brand, T., and Kollmeier, B., 1999b. Entwiklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. *Zeitschrift für Audiologie,* 38(2), 44-56.

Wagener, K., Brand, T., and Kollmeier, B., 1999c. Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. *Zeitschrift für Audiologie,* 38(3), 86-95.

Wagener, K., 2003. *Factors Influencing Sentence Intelligibility in Noise.* Thesis (Dr. Rer. Nat.). Universität Oldenburg.

Wilson, R. H., and McArdle, R., 2005. Speech signals used to evaluate functional status of the auditory system. *Journal of Rehabilitation Research and Development*, 42(4 Suppl 2), 79-94.

Zwicker, E., 1977. Procedure for calculating loudness of temporally variable sounds. *The Journal of the Acoustical Society of America*, 62, 675–682.

Zwicker, E., Fastl, H., and Dallmayr, C., 1984. Basic program for calculating the loudness of sounds from their 1/3 octave band spectra according to ISO532B. *Acustica*, 55, 63-67.

Øygarden, J., 2009. *HiST Taleaudiometri*. Trondheim: HiST.