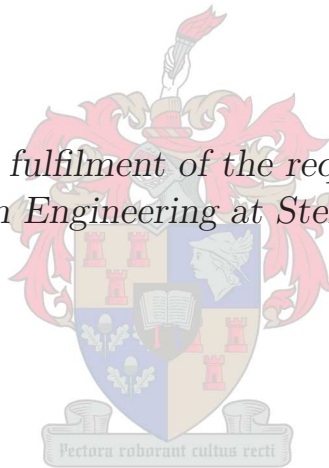


# Automatic Oral Proficiency Assessment of Second Language Speakers of South African English

by  
PIETER F DE V MÜLLER

*Thesis presented in partial fulfilment of the requirements for the degree of  
Master of Science in Engineering at Stellenbosch University*



SUPERVISORS:

Prof. T.R. Niesler

Dr. F. de Wet

Department of Electrical & Electronic Engineering

March 2010

## Declaration

*By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.*

March 2010

COPYRIGHT © 2010 STELLENBOSCH UNIVERSITY

ALL RIGHTS RESERVED

# Abstract

The assessment of oral proficiency forms an important part of learning a second language. However, the manual assessment of oral proficiency is a labour intensive task requiring specific expertise. An automatic assessment system can reduce the cost and workload associated with this task. Although such systems are available, they are typically aimed towards assessing students of American or British English, making them poorly suited for speakers of South African English. Additionally, most research in this field is focussed on the assessment of foreign language students, while we investigate the assessment of second language students. These students can be expected to have more advanced skills in the target language than foreign language speakers.

This thesis presents a number of scoring algorithms for the automatic assessment of oral proficiency. Experiments were conducted on a corpus of responses recorded during an automated oral test. These responses were rated for proficiency by a panel of raters based on five different rating scales. Automatic scoring algorithms were subsequently applied to the same utterances and their correlations with the human ratings determined.

In contrast to the findings of other researchers, posterior likelihood scores were found to be ineffective as an indicator of proficiency for the corpus used in this study. Four different segmentation based algorithms were shown to be moderately correlated with human ratings, while scores based on the accuracy of a repeated prompt were found to be well correlated with human assessments.

Finally, multiple linear regression was used to combine different scoring algorithms to predict human assessments. The correlations between human ratings and these score combinations ranged between 0.52 and 0.90.

# Opsomming

Die assessering van spraakvaardigheid is 'n belangrike komponent van die aanleer van 'n tweede taal. Die praktiese uitvoer van sodanige assessering is egter 'n arbeids-intensiewe taak wat spesifieke kundigheid vereis. Die gebruik van 'n outomatiese stelsel kan die koste en werkslading verbonde aan die assessering van 'n groot aantal studente drasties verminder. Hoewel sulke stelsels beskikbaar is, is dit tipies gemik op die assessering van studente wat Amerikaanse of Britse Engels wil aanleer, en is dus nie geskik vir sprekers van Suid Afrikaanse Engels nie. Verder is die meerderheid navorsing op hierdie gebied gefokus op die assessering van vreemde-taal sprekers, terwyl hierdie tesis die assessering van tweede-taal sprekers ondersoek. Dit is te wagte dat hierdie sprekers se spraakvaardighede meer gevorderd sal wees as dié van vreemde-taal sprekers.

Hierdie tesis behandel 'n aantal evaluasie-algoritmes vir die outomatiese assessering van spraakvaardighede. Die eksperimente is uitgevoer op 'n stel opnames van studente se antwoorde op 'n outomatiese spraaktoets. 'n Paneel van menslike beoordelaars het hierdie opnames geassesseer deur gebruik te maak van vyf verskillende punteskale. Dieselfde opnames is deur die outomatiese evaluasie-algoritmes verwerk, en die korrelasies tussen die beoordelaars se punte en die outomatiese evaluering is bepaal.

In kontras met die bestaande navorsing, is daar gevind dat posterieure waarskynlikheids-algoritmes nie 'n goeie aanduiding van spraakvaardighede gee vir ons datastel nie. Vier algoritmes wat van segmentasies gebruik maak, is ook ondersoek. Die evaluering van hierdie algoritmes het redelike korrelasie getoon met die punte wat deur die beoordelaars toegeken is. Voorts is drie algoritmes ondersoek wat daarop gemik is om die akkuraatheid van herhaalde sinne te bepaal. Die evaluering van hierdie algoritmes het goed gekorreleer met die punte wat deur die beoordelaars toegeken is.

Laastens is liniêre regressie gebruik om verskillende outomatiese evaluering te kombineer en sodoende beoordelaars se punte te voorspel. Die korrelasies tussen hierdie kombinasies en die punte wat deur beoordelaars toegeken is, het gewissel tussen 0.52 en 0.90.

# Acknowledgements

My sincere thanks to:

- My supervisors, Prof. Niesler and Dr. de Wet, for sharing your insight, knowledge, and so much of your time. I could not have wished for mentors more committed beyond the call of duty.
- The NRF, for financial support.
- Wihan & Marike, our coffee breaks will always be a highlight of my time in the lab.
- My parents, Johan & Huibré, for supporting me and giving me so many opportunities.
- Sybil, for your words of motivation and positivity whenever I needed them.
- Jenny, Amanda, Janita and many others, for bringing chocolates, food and words of support to the lab, for listening to me vent my frustration, and for sharing in the joy of my small victories.
- Gert-Jan, for the template this thesis is built on, and also for sharing your passion for all things cool and technical - the linux command line, latex, python, etc.
- Charlene, for the endless kindness, patience and willingness to serve with which you do your job.
- The DSP guys, for dozens of interesting conversations and hundreds of potent cups of coffee.

This project was financially supported by the NRF through a Thuthuka grant for Women in Research awarded to the co-supervisor (Dr. F de Wet) as well as a GUN grant (2072874). Additional funding was also provided by a National HLT Network project, entitled: “Development of resources for intelligent computer-assisted language learning”.

# Contents

<b>Nomenclature</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 System Design . . . . .	2
1.2 Project Background and Thesis Contributions . . . . .	3
1.3 Thesis Structure . . . . .	4
<b>2 Literature Survey</b>	<b>6</b>
2.1 Method of Experimentation . . . . .	6
2.2 Most Relevant Studies . . . . .	7
2.3 Machine Scores . . . . .	8
2.3.1 Segmentation based scores . . . . .	10
2.3.2 HMM likelihood based scores . . . . .	12
2.3.3 Transcription based scores . . . . .	14
2.4 Combination of Machine Scores . . . . .	15
2.4.1 Methods of Combination . . . . .	15
2.4.2 Results . . . . .	17
2.5 Summary and Conclusions . . . . .	18
<b>3 Data Corpus</b>	<b>20</b>
3.1 Test Description . . . . .	20
3.1.1 Test Design . . . . .	21
3.1.2 Test Population . . . . .	21
3.2 Human Ratings . . . . .	22
3.2.1 Rating Scales . . . . .	22
3.2.2 Human Raters . . . . .	22
3.3 Evaluation of Machine Scores . . . . .	27
3.3.1 Averaging Human Ratings . . . . .	27
3.3.2 Spearman's Rank Correlation Coefficient . . . . .	28
3.4 Summary and Conclusions . . . . .	29
<b>4 Automatic Speech Recognition System</b>	<b>30</b>
4.1 Recogniser . . . . .	30

4.1.1	Recognition Output . . . . .	30
4.2	Recognition Strategies . . . . .	31
4.2.1	Finite State Grammar . . . . .	31
4.2.2	Unigram Language Model . . . . .	32
4.2.3	Oracle Finite State Grammar . . . . .	32
4.2.4	Oracle Alignment . . . . .	33
4.2.5	Free Phone Loop Grammar . . . . .	33
<b>5</b>	<b>Posterior Log-Likelihood Scoring</b>	<b>34</b>
5.1	Score Definitions . . . . .	35
5.1.1	All Phones - $GOP_{All}$ . . . . .	36
5.1.2	Only Speech Phones - $GOP_{Speech}$ . . . . .	36
5.1.3	Only Phones in the Context of Speech Phones - $GOP_{Context}$ . . . . .	37
5.1.4	Word Level Normalisation - $GOP_{WordLvl}$ . . . . .	37
5.2	Results . . . . .	37
5.2.1	$GOP_{All}$ . . . . .	38
5.2.2	$GOP_{Speech}$ . . . . .	38
5.2.3	$GOP_{Context}$ . . . . .	39
5.2.4	$GOP_{WordLvl}$ . . . . .	39
5.3	Limiting Phone Scores and Phone Durations . . . . .	40
5.4	Summary and Conclusions . . . . .	45
<b>6</b>	<b>Scores Based On Segmentation</b>	<b>46</b>
6.1	Score Definitions . . . . .	46
6.1.1	<i>Rate of Speech</i> . . . . .	46
6.1.2	<i>Articulation Rate</i> . . . . .	47
6.1.3	<i>Phonation/Time Ratio</i> . . . . .	47
6.1.4	<i>Segment Duration Score</i> . . . . .	47
6.2	Results . . . . .	50
6.2.1	<i>Rate of Speech</i> . . . . .	51
6.2.2	<i>Articulation Rate</i> . . . . .	51
6.2.3	<i>Phonation/Time Ratio</i> . . . . .	51
6.2.4	<i>Segment Duration Score</i> . . . . .	52
6.3	Summary and Conclusions . . . . .	52
<b>7</b>	<b>Scores Based On Repeat Accuracy</b>	<b>53</b>
7.1	Score Definitions . . . . .	53
7.1.1	<i>HResults Accuracy</i> . . . . .	53
7.1.2	<i>HResults Correct</i> . . . . .	53
7.1.3	<i>Weighted Correct</i> . . . . .	54
7.2	Results . . . . .	55

7.2.1	<i>HResults Accuracy</i> . . . . .	55
7.2.2	<i>HResults Correct</i> . . . . .	56
7.2.3	<i>Weighted Correct</i> . . . . .	56
7.3	Summary and Conclusions . . . . .	57
<b>8</b>	<b>Combination of Scores</b>	<b>58</b>
8.1	Linear Regression . . . . .	58
8.2	Application of MLR to Scores, Ratings and Marks . . . . .	60
8.3	Evaluation . . . . .	63
8.3.1	<i>Hesitation</i> . . . . .	64
8.3.2	<i>Pronunciation</i> . . . . .	64
8.3.3	<i>Intonation</i> . . . . .	65
8.3.4	<i>Success</i> . . . . .	65
8.3.5	<i>Accuracy</i> . . . . .	66
8.3.6	Oral Mark . . . . .	67
8.3.7	Progress Mark . . . . .	68
8.4	Summary and Conclusions . . . . .	69
<b>9</b>	<b>Summary and Conclusions</b>	<b>71</b>
9.1	Human Ratings . . . . .	71
9.2	Machine Score Algorithms . . . . .	71
9.3	Combination of Machine Scores . . . . .	73
9.4	Recommendations for Future Research . . . . .	73
	<b>Bibliography</b>	<b>75</b>
	<b>A Reading Task Prompts</b>	<b>78</b>
	<b>B Repeating Task Prompts</b>	<b>79</b>
	<b>C Distributions of Monophone Durations</b>	<b>80</b>
	<b>D Inter-Score Correlations</b>	<b>84</b>
	<b>E Automated Oral Test Software</b>	<b>86</b>



# List of Figures

1.1	Diagram showing the implementation and design of an automatic assessment system. . . . .	3
3.1	Reading task rating scales used to assess (a) degree of <i>Hesitation</i> , (b) <i>Pronunciation</i> and (c) <i>Intonation</i> . . . . .	23
3.2	Repeating task rating scales used to assess (a) degree of <i>Success</i> and (b) <i>Accuracy</i> . . . . .	23
3.3	Mean ratings assigned for each of the (a) reading and (b) repeating task rating scales. Horizontal bars show the standard deviations. . . . .	24
3.4	Average inter-rater agreement for each of the (a) reading and (b) repeating task rating scales. . . . .	25
4.1	Example of a finite state grammar network for the hypothetical sentence “Close the door.” . . . .	32
4.2	Network showing the structure of the unigram LM recognition strategy. . . . .	33
5.1	Mismatched segmentation between phones selected by forced alignment and those selected by free phone loop recognition. . . . .	36
5.2	Distribution of phone level GOP scores assigned to phones in speech context for the reading task. . . . .	41
5.3	Distribution of durations of phones in speech context for the reading task. . . . .	42
5.4	Correlations of <i>Pronunciation</i> ratings with $GOP_{Context}$ scores against allowed maximum phone duration. . . . .	43
5.5	<i>Subset A</i> : Correlations of <i>Pronunciation</i> ratings with $GOP_{Context}$ scores against allowed maximum phone duration. . . . .	44
5.6	<i>Subset B</i> : Correlations of <i>Pronunciation</i> ratings with $GOP_{Context}$ scores against allowed maximum phone duration. . . . .	44
6.1	Histogram of normalised durations of the phone “ <i>sw</i> ” based on training data. . . . .	49
6.2	Histogram of normalised durations of the phone “ <i>sw</i> ” after smoothing with median filter. . . . .	49
6.3	Discrete probability distribution of normalised duration of the phone “ <i>sw</i> ” based on training data. . . . .	50

8.1	Hypothetical example of simple linear regression. . . . .	59
C.1	Discrete probability distributions of the normalised duration, $f(q)$ , of monophones. . . . .	80
C.2	Discrete probability distributions of the normalised duration, $f(q)$ , of monophones. . . . .	81
C.3	Discrete probability distributions of the normalised duration, $f(q)$ , of monophones. . . . .	82
C.4	Discrete probability distributions of the normalised duration, $f(q)$ , of monophones. . . . .	83
E.1	Startup window of automated oral test software. . . . .	86
E.2	Dialogue window of automated oral test software. . . . .	87

# List of Tables

2.1	Summary of correlations between human ratings and machine scores in a number of different studies. . . . .	9
2.2	Performance of machine score combination methods relative to the correlation of posterior HMM-LL with human ratings. . . . .	18
3.1	Intra-rater correlations for human raters. . . . .	25
3.2	Correlations between rating scales and students' academic marks. . . . .	26
3.3	Example of ranks for calculating Spearman's rank correlation coefficient. . . . .	28
5.1	Correlation of $GOP_{All}$ scores with human ratings for different rating scales and recognition strategies. . . . .	38
5.2	Correlation of $GOP_{Speech}$ scores with human ratings for different rating scales and recognition strategies. . . . .	39
5.3	Correlation of $GOP_{Context}$ scores with human ratings for different rating scales and recognition strategies. . . . .	39
5.4	Correlation of $GOP_{WordLvl}$ scores with human ratings for different rating scales and recognition strategies. . . . .	40
6.1	Correlation of <i>Rate of Speech</i> scores with human ratings for different rating scales and recognition strategies. . . . .	51
6.2	Correlation of <i>Articulation Rate</i> scores with human ratings for different rating scales and recognition strategies. . . . .	51
6.3	Correlation of <i>Phonation/Time Ratio</i> scores with human ratings for different rating scales and recognition strategies. . . . .	52
6.4	Correlation of <i>Segment Duration Score</i> scores with human ratings for different rating scales and recognition strategies. . . . .	52
7.1	Four different weight sets associated with the <i>Weighted Correct</i> ranks. . . . .	55
7.2	Correlation of <i>HResults Accuracy</i> scores with human ratings for different rating scales and recognition strategies. . . . .	56
7.3	Correlation of <i>HResults Correct</i> scores with human ratings for different rating scales and recognition strategies. . . . .	56
7.4	Correlation of <i>Weighted Correct</i> scores with human ratings for different rating scales, recognition strategies and weight sets. . . . .	57

8.1	Different configurations of target and predictor variables for MLR. . . . .	60
8.2	Categories of machine scores, human ratings and academic marks. . . . .	61
8.3	Descriptions of machine scores listed in Table 8.2. . . . .	62
8.4	Predictor set consisting of machine scores for the reading and repeating tasks after trimming strongly correlated scores. . . . .	63
8.5	Results for MLR predictions of <i>Hesitation</i> ratings based on reading task machine scores. . . . .	64
8.6	Results for MLR predictions of <i>Pronunciation</i> ratings based on reading task machine scores. . . . .	65
8.7	Results for MLR predictions of <i>Intonation</i> ratings based on reading task machine scores. . . . .	65
8.8	Results for MLR predictions of <i>Success</i> ratings based on repeating task machine scores. . . . .	66
8.9	Results for MLR predictions of <i>Accuracy</i> ratings based on repeating task machine scores. . . . .	66
8.10	Results for MLR predictions of oral marks based on human proficiency ratings. . . . .	67
8.11	Results for MLR predictions of oral marks based on machine scores. . . . .	67
8.12	Results for MLR predictions of progress marks based on human proficiency ratings. . . . .	68
8.13	Results for MLR predictions of progress marks based on machine scores. . . . .	68
9.1	Summary of the correlations between machine scores and human ratings. . . . .	72
D.1	Inter-score correlations for the reading task. . . . .	84
D.2	Inter-score correlations for the repeating task. . . . .	85

# Nomenclature

$Acc_{HResults}$	HResults Accuracy
ART	Articulation Rate
ASR	Automatic Speech Recognition
AST	African Speech Technology
CALL	Computer Assisted Language Learning
$Cor_{HResults}$	HResults Correct
$Cor_{Weighted}$	Weighted Correct
EBNF	Extended Backus-Naur Form
FSG	Finite State Grammar
GOP	Goodness of Pronunciation
HMM	Hidden Markov Model
HMM-LL	Hidden Markov Model Log-Likelihood
HTK	Hidden Markov Model Toolkit
L2	Target Language
LM	Language Model
MFCC	Mel-Frequency Cepstral Coefficient
MLR	Multiple Linear Regression
PTR	Phonation/Time Ratio
ROS	Rate of Speech
RSS	Residual Sum of Squares
SAE	South African English
SDS	Segment Duration Score
SLaTE	Speech and Language Technology in Education
SLR	Simple Linear Regression
WEKA	Waikato Environment for Knowledge Analysis

# Chapter 1

## Introduction

It is often said that the world is getting smaller. International travel is becoming less expensive and many company structures span international borders. Along with advances in telecommunication technology and the expansion of the internet, this means people are encountering foreign languages more often. It seems likely that acquiring a second language will be a common need amongst citizens of the emerging “global village”.

Part of learning to speak a second language is the assessment of oral proficiency. It allows the student to receive constructive feedback regarding systematic mistakes, or to seek instruction suitable to his level of proficiency. Also, people seeking employment or wishing to immigrate often require endorsements of their oral proficiency in a specified language. However, manual assessment of oral proficiency is a labour intensive task that requires specific expertise. This makes automatic assessment of oral proficiency an attractive option. This is perhaps especially true in the developing world, where the number of students per teacher is often high and expertise in short supply.

The research presented in this thesis forms part of an ongoing effort to develop a system capable of automatically assessing the oral proficiency of large numbers of students in the specific context of the Stellenbosch University Education Faculty. Students at the Faculty are required to obtain a language endorsement on their teaching qualification. English language modules are offered to develop the students’ English skills so as to enable them to either teach their subjects in English (the higher endorsement), or to use English in professional communication (the lower endorsement). Students need to select an English language module which is appropriate for their language skill level, making it necessary to assess their oral proficiency before enrolment and to monitor their progress regularly thereafter. With between 100 and 200 students per staff member, the current system relies heavily on multiple choice reading and writing tests, since the labour intensive assessment of oral skills is not a feasible option. However, students regard oral proficiency as an important component of their teaching skills and are not satisfied with the current tests.

A project was subsequently started to develop an automatic oral proficiency assessment system. The system is intended to reduce the workload associated with proficiency assessments, allow speedy availability of results to students, and be more objective than human

assessments, which are often very subjective. There are commercial products with similar functionality, such as Versant<sup>1</sup>, EyeSpeak<sup>2</sup>, Carnegie Speech Assessment<sup>3</sup> and EduSpeak<sup>4</sup>. However, these products are expensive, and the speech recognisers they employ are focussed on students of British or American English, making them poorly suited for speakers of South African English. Additionally, these products are aimed at students of English as a *foreign language*, which implies a substantial contrast between high and low oral proficiency. The students at Stellenbosch University are predominantly *second language* speakers, whose proficiency in English ranges from intermediate to advanced. Because of this difference in proficiency range, the same automatic assessment approach used for foreign language speakers may not apply directly to second language speakers. The difference between foreign and second language speakers is defined further in Section 2.5.

## 1.1 System Design

Figure 1.1 shows a diagram of the automatic assessment system described in this thesis. The left branch represents the structure a completed system would have, while the human ratings branch on the right is only required while the system is being developed.

The following processes are defined:

**Oral Test.** An oral test is used to collect utterances from the **test population**. The **test design** determines which tasks the students must perform. The test and test population are described in Chapter 3.

**Speech Recogniser.** Automatic speech recognition is performed on the **recorded utterances** collected during the oral test. **Scoring algorithms** utilising the features extracted during recognition are used to automatically calculate **machine scores** for the utterances. The automatic speech recognition process is described in Chapter 4. The scoring algorithms are described in Chapters 5, 6 and 7.

**Human Raters.** While developing the automatic assessment system, human raters are asked to rate the **recorded utterances** for proficiency using **rating scales**. The resulting **human ratings** are compared to the **machine scores** to evaluate the latter's potential for predicting a human rater's assessment of a test utterance. The human ratings and rating scales are discussed in Chapter 3. Comparisons between machine scores and human ratings are presented in Chapters 5, 6 and 7.

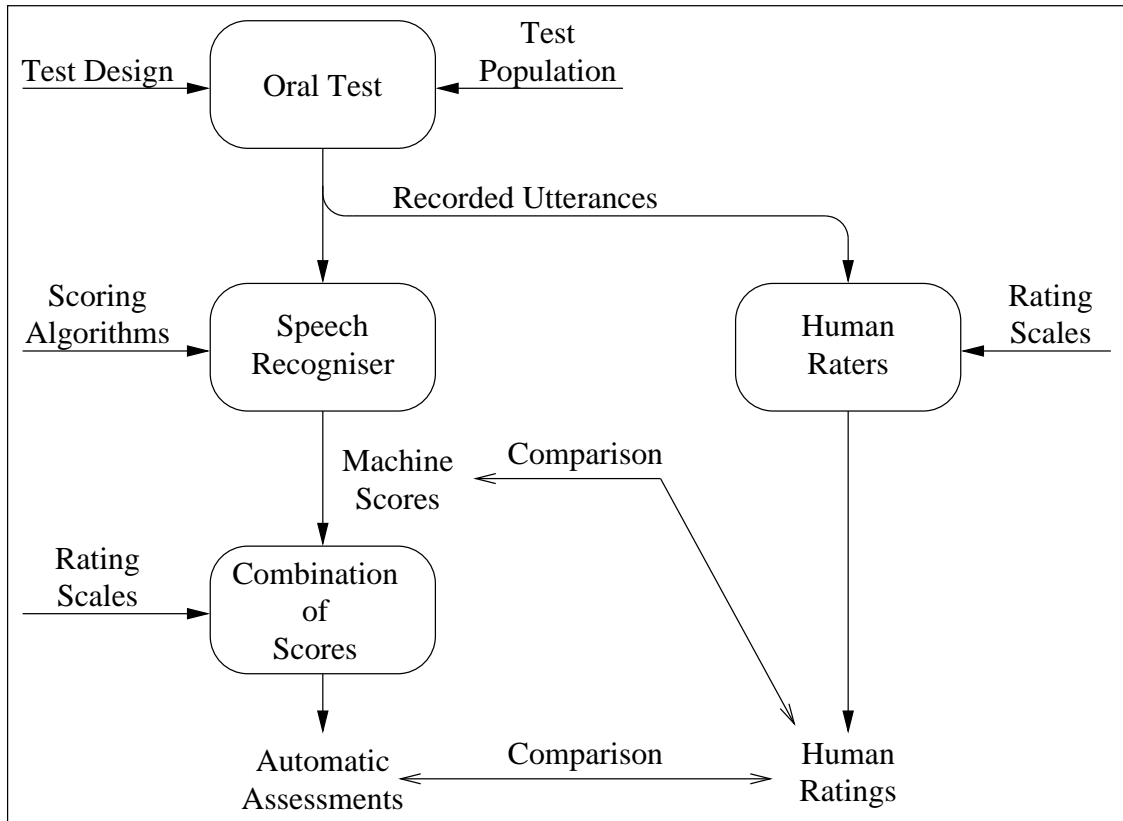
---

<sup>1</sup>[www.ordinate.com](http://www.ordinate.com)

<sup>2</sup>[www.eyespeakenglish.com](http://www.eyespeakenglish.com)

<sup>3</sup>[www.carnegiespeech.com](http://www.carnegiespeech.com)

<sup>4</sup>[www.eduspeak.com](http://www.eduspeak.com)



**Figure 1.1:** Diagram showing the implementation and design of an automatic assessment system.

**Combination of Scores.** Multiple **machine scores** can be combined to determine **automatic assessments** of students. In Chapter 8, we present the use of multiple linear regression to predict **human ratings**. These predicted ratings are then compared to the actual human ratings to evaluate their accuracy. We evaluate the quality of the predictions in terms of the correlation between the predicted values and the human ratings.

## 1.2 Project Background and Thesis Contributions

In 2005, staff at the Stellenbosch University Faculty of Education expressed the desire to assess the oral proficiency of large numbers of students automatically. An automated telephonic oral test was consequently developed and 30 students took part in a pilot study.

In 2006, a larger group of students took the test and their recorded responses were manually rated for proficiency, based on four Likert scales. The ratings were subsequently compared with the *rate of speech* of these responses, determined by an automatic speech recogniser. This experiment is described in [1] and [2].

The recorded utterances were re-evaluated in 2007 using a revised set of rating scales.



Together with the recorded test responses, these human ratings compose the corpus of data used for the research in this thesis. Three automatic scoring algorithms were applied to the recorded utterances, and these scores were compared with the manually assigned proficiency ratings. The results of this study are presented in [3].

This thesis describes the contributions to the project by the author during 2008 and 2009. A number of additional automatic scoring algorithms are evaluated on the test data to determine their potential for assessing oral proficiency in the context of this project. For some scoring algorithms, an attempt is made to improve the results obtained during earlier stages of the project. Finally, an effort is made to combine different automatic scoring algorithms to create assessments of proficiency that resemble those determined manually by human raters. All experiments described in this thesis were performed by the author, except where explicitly indicated otherwise. The oral test recordings and associated human ratings were pre-existing, however.

The work described in this thesis has led to two published papers, [4] and [5], which the author presented at the associated conferences. The presentation of [5] was awarded with the prize for best student presentation at SLATE 2009, an international event.

## 1.3 Thesis Structure

**Chapter 2 - Literature Survey.** This chapter provides an overview of relevant previous research in the field of automatic oral proficiency assessment. The method of experimentation is described and a number of scoring algorithms are introduced. We also examine different methods of combining automatic scoring algorithms.

**Chapter 3 - Data Corpus.** This chapter describes the corpus of data used for the research conducted during this thesis. The design and implementation of the automated oral test is presented, along with the manual rating process. Finally, we discuss the method used to evaluate the performance of automatic scoring algorithms.

**Chapter 4 - Automatic Speech Recognition System.** All the automatic scoring algorithms presented in this thesis depend on automatic speech recognition of the utterances to be assessed. This chapter describes the recogniser and recognition strategies used to calculate proficiency scores.

**Chapter 5 - Posterior Log-Likelihood Scoring.** This chapter presents the *Goodness of Pronunciation* scoring algorithm and variations thereof. We compare the scores with manually obtained proficiency ratings and investigate possible ways of improving the performance of posterior likelihood scoring.

**Chapter 6 - Scores Based On Segmentation.** This chapter presents four scoring algorithms based on the phonetic segmentation of the utterances to be assessed. The scores are the *Rate of Speech*, the *Articulation Rate*, the *Phonation/Time Ratio* and

the *Segment Duration Score*. We evaluate the algorithms by comparing the scores with manually obtained proficiency ratings.

**Chapter 7 - Scores Based On Repeat Accuracy.** This chapter presents three scoring algorithms based on the accuracy of a repeated utterance. The scores are *HResults Accuracy*, *HResults Correct* and the *Weighted Correct*. As before, we evaluate the algorithms by comparing the scores with manually obtained proficiency ratings.

**Chapter 8 - Combination of Scores.** In this chapter we investigate the combination of different scoring algorithms using multiple linear regression. An introduction to linear regression is provided. Scores are combined to predict ratings from each manual rating scale separately, allowing us to identify which scoring algorithms are effective predictors of which aspects of oral proficiency.

**Chapter 9 - Summary and Conclusions.** This chapter provides a summary of the thesis and the conclusions reached. Recommendations for future research are also provided.

# Chapter 2

## Literature Survey

This chapter presents an overview of existing research on automated oral proficiency assessment. The majority of the work in this field relates to computer assisted language learning (CALL) applications for foreign language speakers. In some cases the aim is to assess the overall oral proficiency of the students, while other studies aim to identify mispronounced words or phones, in order to give constructive feedback. Although studies vary in the scales used to assess proficiency, there is significant overlap in the machine scoring algorithms applied.

We begin the chapter with a description of the method of experimentation shared by many of the studies presented here. Next, we give an overview of the relevant studies, focussing on the composition of each group's data corpus. We subsequently describe the machine scoring algorithms used, a number of which will be investigated in this thesis. Finally, we present methods of combining machine scores to better assess oral proficiency automatically.

### 2.1 Method of Experimentation

When carrying out experiments in automated oral language proficiency assessment, Witt et al. [6], Neumeyer et al. [7], Cucchiaroni et al. [8] and Hacker et al. [9], all use a similar approach. A speech recogniser is trained using recordings of native speakers of the language under study. A set of utterances by the target test group, usually consisting of second language speakers, is then recorded. These utterances are rated by a panel of evaluators, producing what are known as the *human ratings*. The same utterances are then processed automatically by the speech recogniser, extracting a set of objective or quantitative features commonly referred to as the *machine scores*. Finally, correlations between the *human ratings* and the *machine scores* are determined to identify those features that can be used as effective predictors of the ratings assigned by human evaluators. Franco et al. [10] and Cincarek et al. [11] go a step further by considering various methods to combine different machine scores, in some cases increasing correlation with the assigned human ratings.

## 2.2 Most Relevant Studies

Four recent studies are summarised briefly in the following, since they have been found to be of direct and important relevance to the research presented in this thesis. Only the structure of the experimental work is discussed, while mathematical detail is described later in Sections 2.3 and 2.4.

### Witt & Young

Witt & Young, [6], set out to measure pronunciation quality at the phone level using a posterior log-likelihood machine score which they term the *Goodness of Pronunciation (GOP)*. Their experiments make use of ten students of English as a second language, each of whom read 120 sentences. These students had different mother-tongues. Six raters were asked to annotate the recorded utterances, marking mispronunciations. A subset of these sentences was marked by all raters. This common set of annotations was used to compare the assessments of the raters based on four performance measures: *Strictness*, *agreement*, *cross-correlation* and *overall phone correlation*. *Strictness* is defined as the fraction of phones that were marked as mispronunciations (*rejected*). *Agreement* is an indication of how similar two annotations are, taking all phones into account. The *cross-correlation* determines the agreement of rejections between two transcriptions, while the *overall phone correlation* compares the rejection statistics for each phone between two transcriptions. The same measures were used to evaluate the performance of the Goodness of Pronunciation scores.

### Neumeyer et al. and Franco et al.

The work by Neumeyer et al. [7] studies the correlations between a number of machine scores and human ratings for *fluency*. As data for the experiment, 100 American students of French read about 30 sentences each from newspapers. Ten raters were asked to rate a subset of this non-native data, allowing inter- and intra-rater reliability to be calculated. Only the five most reliable raters were asked to rate the entire data set. A number of machine scores and their correlations with the human ratings were then calculated.

Franco et al. [10] continued the above experiment by considering various methods of combining machine scores in an effort to achieve higher correlations with human ratings.

### Cucchiaroni et al.

The study by Cucchiaroni et al. consisted of three phases. The first phase, [8], focused on the reliability of human raters. A set of 80 speakers of Dutch with varied proficiency levels each read ten sentences over the telephone. Three separate groups of raters were then tasked with rating these utterances in terms of *overall pronunciation quality*, *segmental quality*, *fluency* and *speech rate*. The raters did not receive any specific instructions on how to use the rating scales. One group consisted of three phoneticians, the other two groups consisted of three

speech therapists each. Each group rated the entire data set, with some overlap between individual raters for comparative purposes. Furthermore, some material was presented to each rater twice, to assess consistency. After some normalisation, the study found good inter- and intra-rater correlations and concluded that all raters involved in the study rated the material in a similar way.

The second phase of the study, [12], focused on the use of machine scores for the automatic rating of read speech. The material and ratings obtained during the first phase of the experiment was used, and the ratings correlated with a number of different machine scores.

In the third phase, [13], the authors applied the previously studied machine scores to spontaneous speech. The spontaneous speech material was recorded in a language laboratory and consisted of two sets of recordings. One set consisted of intermediate level speakers answering questions and motivating their answers in utterances of 30 seconds each. The second set consisted of beginner level speakers answering simple questions in 15 second utterances. The material was rated by teachers of Dutch as a second language, with no overlap of material between raters. Machine scores were calculated from the spontaneous speech recordings, correlations with human ratings calculated, and the results compared with those previously found for read speech. The study showed that automatic rating is more effective when applied to read speech than when applied to spontaneous speech, although the many differences between the two experiments made comparison difficult.

### **Hacker et al.**

Hacker et al. calculated a large number of machine scores for two existing databases of non-native speech, as well as the correlations of these scores with human ratings [9]. One database used was the ATR/SLT non-native database, for which 96 speakers with various mother-tongues each read 48 English sentences. The utterances in this database were rated by 15 English teachers, who assigned a rating based on pronunciation and fluency to each sentence. The other database used was the PF-STAR non-native database, made up of read sentences by young children with various mother-tongues. We will concentrate on the results for the ATR/SLT database.

Cincarek et al. extended this study by combining machine scores to classify words as correctly pronounced or mispronounced [11].

## **2.3 Machine Scores**

In [7], Neumeyer et al. describe the system used to determine machine scores for a given speech waveform. The waveform is converted into a sequence of mel-frequency cepstral coefficients (MFCC) for use by a speech recogniser. The recogniser then divides the audio into segments based on the start and end times of different phones, using a human transcription of the utterance and forced Viterbi alignment. A number of machine scores can be calculated based on this segmentation. Probabilities calculated by the speech recogniser during the

Viterbi alignment allow the calculation of machine scores based on a hidden Markov model (HMM) likelihood. Other machine scores can be calculated using the transcription of the utterance and language specific features.

This section describes a number of machine scores and their correlations with human ratings, as determined in the studies introduced in Section 2.2 and a previous stage of the research presented in this thesis [2]. These correlations are summarised in Table 2.1.

Whenever a correlation value is given in this chapter, its absolute value is used. The sign of a correlation value depends on the nature of the machine score and the definition of the human rating scale the score is being correlated with, making the sign unimportant when comparing correlations between studies, since not all authors define their rating scales in a similar manner.

Machine Score	Witt & Young	Neumeyer et al.	Cucchiari et al. (Read)	Cucchiari et al. (Spontaneous, Beginner Level)	Cucchiari et al. (Spontaneous, Intermediate Level)	Hacker et al.	De Wet et al.
Total Duration			0.92				
Rate of Speech			0.92	0.57	0.39	0.39	0.58
Articulation Rate			0.83	0.07	0.05		
Phonation/Time Ratio			0.86	0.46	0.39		
Segment Duration Score		0.86				0.46	
Syllabic Timing		0.73					
Number of Silent Pauses			0.84	0.33	0.49	0.32	
Total Duration of Pauses			0.84	0.45	0.40	0.33	
Mean Length of Pauses			0.53	0.08	0.01		
Mean Length of Runs			0.85	0.49	0.65		
Number of Filled Pauses			0.25	0.21	0.21		
Number of Dysfluencies			0.15	0.07	0.27		
Average HMM-LL		0.48				0.42	
Posterior HMM-LL	0.72	0.84	0.62			0.52	
Recognition Accuracy		0.47				0.45	
PhoneSeq						0.40	

**Table 2.1:** Summary of correlations between human ratings and machine scores in a number of different studies.

### 2.3.1 Segmentation based scores

The scores described here are derived from the segmentation of an utterance into its constituent phones. Segmentation can be done manually or automatically with the Viterbi algorithm.

#### Total Duration

Cucchiaroni et al. [8], calculated the correlation of *total utterance duration*,  $T_{Total}$ , with human ratings.  $T_{Total}$  is the duration of an utterance in seconds or number of frames. In the experiment by Cucchiaroni et al. all speakers read the same prompts, therefore all utterances contained the same number of phones, making comparison of total utterance duration possible. After normalising the human ratings,  $T_{Total}$  had a correlation of 0.92 with the human ratings for *fluency*.

#### Rate of Speech

Cucchiaroni et al. defined *rate of speech (ROS)* as  $\frac{\text{Number of Phones}}{T_{Total}}$ . In the first phase of the study by Cucchiaroni et al. [8], *ROS* had a correlation of 0.81 with the human ratings for *overall pronunciation*, better than that achieved by either *total duration* and *posterior HMM log-likelihood* (see section 2.3.2). In the second phase of the study [12], *ROS* presented a correlation of 0.92 with the normalised human ratings for *fluency*, a higher correlation than any other machine score investigated in that experiment.

The third phase of the study by Cucchiaroni et al. [13], investigated the correlation of *ROS* with fluency ratings for spontaneous speech. Of the machine scores calculated, *ROS* presented the best correlation (0.57) with the human ratings for the beginner level speakers, but did not present significant correlation with human ratings for the intermediate level speakers. In general, correlations calculated for spontaneous speech were significantly lower than those calculated for read speech.

In the study by Hacker et al. [9], *ROS* is also among the machine scores calculated. A *ROS* score based on the number of words in an utterance as well as the usual *ROS* based on the number of phones was calculated, along with the reciprocals of both. Of these four scores, the reciprocal of the phone-based *ROS* had the highest correlation, 0.39, with the human ratings for “*pronunciation and fluency*”. The best correlation achieved in the study was 0.52, for a normalised form of the posterior HMM log-likelihood (see section 2.3.2).

In an earlier phase of the research presented in this thesis, De Wet et al. [2], calculated *ROS* for read, repeated (after a prompt) and spontaneous speech by proficient second language students of English. Correlation with human ratings for *pronunciation* varied between 0.48 (spontaneous speech) and 0.58 (repeated speech). The utterances were both automatically and manually transcribed. Correlations between the *ROS* values calculated from manual transcriptions and the *ROS* values calculated from automatic transcriptions varied between 0.86 (spontaneous speech) and 0.98 (read speech). This shows that although

automatic transcriptions are not perfect, the *ROS* values based on such transcriptions are quite reliable.

### Articulation Rate

In the study by Cucchiari et al. [13], the authors determined  $T_{NoPause}$ , the duration of an utterance without internal pauses, where a pause is defined as silence of at least 0.2 seconds. This allowed the calculation of the *articulation rate*, defined as  $\frac{\text{Number of Phones}}{T_{NoPause}}$ . The articulation rate had a correlation of 0.83 with the normalised human fluency ratings for read speech [12]. However, the articulation rate showed a weak correlation with the human ratings for spontaneous speech. This is attributed to the high number of pauses that occur naturally in spontaneous speech and the fact that these pauses penalise the articulation rate.

### Phonation/Time Ratio

Cucchiari et al. defined the *phonation/time ratio* (*PTR*) for an utterance as  $\frac{T_{NoPause}}{T_{Total}} \times 100\%$ . For read speech [12], *PTR* had a correlation of 0.86 with normalised human ratings for *fluency*, where the best correlation was that with *ROS*, 0.92. For spontaneous speech [13], the correlation with human ratings was 0.46 for the beginner level group and 0.39 for the intermediate level group.

### Segment Duration Score

In the study by Neumeyer et al. [7] the *segment duration score* (*SDS*) was calculated by comparing the duration of a segment from Viterbi alignment,  $d_i$ , with the duration expected for that particular phone based on native training data. The argument is that for less proficient speakers, thinking about how to pronounce a particular phone will result in phone durations that differ from those that may be expected for native speakers.

The duration must be normalised for the speaker’s rate of speech:

$$f(q_i) = d_i \cdot \text{ROS}$$

where  $f(q_i)$  is the normalised duration of phone  $q_i$ . The *SDS* is then calculated as the log-probability of the normalised segment duration, using a discrete distribution of durations for the particular phone gathered from native training data. These log-probabilities are averaged over all segments in the utterance to be rated:

$$SDS = \frac{1}{M} \sum_{i=1}^M \log \left( p(f(q_i)|q_i) \right)$$

where  $M$  is the number of segments and  $q_i$  is the phone that corresponds to the  $i^{th}$  segment.

In the study by Neumeyer et al. [7], the *SDS* was computed for each non-native speaker and averaged over 30 sentences. Phones in the context of silence were disregarded. The



*SDS* had a correlation of 0.86 with the human ratings for *pronunciation*, the highest of the machine scores investigated in that experiment.

In the study by Hacker et al. [9], a similar score named *DurationScore* had a correlation of 0.46 with human ratings. The same authors also calculated another measure based on the expected duration of phones, called *DurationLUT*. The deviation  $|d_i - d_{q_i}|$  was determined, where  $d_{q_i}$  is the average duration of the corresponding phone for segment  $i$  based on native training data. Correlations of 0.30 and 0.28 were calculated for the mean and the variance of this deviation respectively.

### Syllabic Timing

Neumeyer et al. [7] propose *syllabic timing* as a proficiency measure based on the tendency of non-native speakers to impose their native tongue’s rhythm on the second language. The time duration between the centres of vowels in an utterance are measured based on the Viterbi alignment, and then normalised. From a distribution of these durations, a syllabic timing score is calculated. The authors argue that syllabic timing is a more robust measure than *ROS*, as any speech-like signal of the right duration could produce high *ROS* scores. Syllabic timing had a correlation of 0.73 with human ratings.

### Scores based on Hesitation Phenomena, Pauses and Runs

Cucchiaroni et al. [13] manually transcribed utterances using symbols for *pauses* (defined as a silence of at least 0.2 seconds), *filled pauses*, and different types of *noise*. Repetitions, restarts and repairs, grouped as *hesitation phenomena* or *dysfluencies*, were transcribed exactly as they were pronounced. These transcriptions allowed the calculation of a number of machine scores based on a speaker’s *pauses*, *hesitations* and *runs* (uninterrupted speech between pauses). Hacker et al. [9] also considered two of these features, by calculating the *number of silent pauses* and the *total duration of pauses*.

The correlations with human ratings for both studies are given in Table 2.1. Note that the scores *number of filled pauses* and *number of dysfluencies* can not currently be calculated automatically, as manual transcriptions of the material to be scored are required.

### 2.3.2 HMM likelihood based scores

When processing an utterance, a speech recogniser can output probabilities showing the certainty with which a phone has been identified. These probabilities, based on the match between the audio signal and the given phone’s HMM, can be used to calculate a number of scores based on the *HMM likelihood*.

#### Average HMM Log-Likelihood

Probably the most basic HMM likelihood based scores are the *global-* and *local average HMM log-likelihoods*, (*HMM-LL*), investigated by Neumeyer et al. [7]. These scores are based on

the logarithm of the likelihood of the most probable path found by the Viterbi algorithm during phone segmentation of an utterance. The HMM-LL for the acoustic segment  $O_i$  consisting of  $N_i$  frames aligned with phone  $q_i$  chosen by the Viterbi algorithm, is:

$$\log(p(O_i|q_i)) = \log\left(\prod_{n=1}^{N_i} p(s_{i_n}|s_{i_{n-1}})p(o_{i_n}|s_{i_n})\right)$$

where  $o_{i_n}$  denotes the acoustic observation corresponding to the  $n_{th}$  frame of the segment  $O_i$ ,  $s_{i_n}$  the HMM state aligned with this observation, and  $p(s_{i_n}|s_{i_{n-1}})$  the HMM transition probability between states  $s_{i_n}$  and  $s_{i_{n-1}}$ . The automatic speech recognition process is discussed in more detail in Section 4.1.1.

When summing the HMM-LL scores over all acoustic segments in a sentence, the total must be normalised for the length of the sentence. Two methods to achieve this have been proposed. The *global average HMM-LL* score  $G$  is defined as the sum of all  $M$  segment HMM-LL scores in an utterance normalised by its total duration:

$$G = \frac{\sum_{i=1}^M \log(p(O_i|q_i))}{\sum_{i=1}^M d_i}$$

where  $d_i$  is the duration of the  $i^{th}$  segment, often expressed as the number of frames,  $N_i$ . A possible disadvantage of the global average HMM-LL score is that it is dominated by longer phones, while shorter phones may have a more important perceptual effect. As compensation for this effect, the *local average HMM-LL* score  $L$  has been suggested, where the score for each segment is normalised by its duration before summation over all the segments of the sentence:

$$L = \frac{1}{M} \sum_{i=1}^M \frac{\log(p(O_i|q_i))}{d_i}$$

In the study by Neumeyer et al. [7], the global average HMM-LL scores had a correlation of 0.31 with human ratings, while the local average HMM-LL scores had a correlation of 0.48 with human ratings. Hacker et al. [9], calculated correlations for a number of variations on the average HMM-LL. By normalising the local average HMM-LL score with *ROS*, a correlation of 0.42 was achieved. Replacing the phone duration  $d_i$  with the statistically predicted phone duration from a duration statistic look-up-table lead to a correlation 0.43.

### Posterior HMM Log-Likelihood

A number of authors investigate the correlation between the log of the *posterior* HMM-likelihood and human ratings for fluency. Witt and Young [6], propose the *Goodness of Pronunciation (GOP)* score to identify individual mispronounced phones based on a rejection threshold. Neumeyer et al. [7] propose essentially the same measure, but refer to it as the *Log Posterior Score*, calculated per frame and averaged to produce a sentence pronunciation score. Cucchiarini et al. [8] present the *Likelihood Ratio* and Hacker et al. [9] the *LikeliRatio*,

both based on the difference between the log-likelihood resulting from forced alignment and the log-likelihood resulting from unconstrained phone loop recognition.

The *GOP* score for a phone in an utterance is defined as the duration normalised log of the posterior likelihood  $P(q_i|O_i)$  that the speaker uttered phone  $q_i$  given the acoustic segment  $O_i$ .

$$GOP(q_i) \equiv |\log(P(q_i|O_i))|/N_i$$

Bayes' rule gives

$$GOP(q_i) = \left| \log \left( \frac{p(O_i|q_i)P(q_i)}{\sum_{j=1}^J p(O_i|q_j)P(q_j)} \right) \right| / N_i,$$

where  $J$  is the total number of phone models. When assuming that all phones are equally likely and that the sum in the denominator can be approximated by its maximum, the *GOP* score is given by

$$GOP(q_i) \approx \left| \log \left( \frac{p(O_i|q_i)}{\max_{j=1}^J p(O_i|q_j)} \right) \right| / N_i \quad (2.1)$$

This is equivalent to the log of the ratio between the likelihood of the phone chosen by a forced alignment and the likelihood of the most likely phone, as determined by using a free phone loop. The *Likelihood Ratio* used by Cucchiarini et al. as well as the *LikeliRatio* used by Hacker et al. are defined in this way.

In the study by Witt and Young [6], the *GOP* score was calculated for each phone and the phone marked as mispronounced if the *GOP* score fell below a certain rejection threshold. The basic *GOP* method resulted in a cross-correlation of 0.62 with human rater phone rejections. A number of refinements improved the cross-correlation to 0.72.

In the study by Neumeyer et al. [7], the authors report a correlation of 0.84 between the log posterior score and human ratings, while Cucchiarini et al. report correlation values between 0.55 and 0.68. Hacker et al. [9] report correlations of between 0.48 and 0.52 between their *LikeliRatio* and human ratings, using various methods of normalisation.

### 2.3.3 Transcription based scores

The scores described here depend on the transcription of the utterance to be rated or on the specific language being used.

#### Recognition Accuracy

Neumeyer et al. [7] as well as Hacker et al. [9] investigate the correlation between human ratings and a score based on the phone recognition accuracy of an automatic speech recogniser. The authors argue that a recogniser trained on native data will be prone to reject phones pronounced in a non-native manner. Neumeyer et al. report a correlation with human ratings of 0.47, while Hacker et al. report 0.45.

## PhoneSeq

Hacker et al. [9] estimate a phone bigram language model (LM) on native data. This allows the a priori probability  $\log P(\mathbf{q}|\text{LM})$  of an observed phone sequence  $\mathbf{q}$  to be calculated. Normalisation with *ROS* results in a correlation of 0.40 with human ratings.

## 2.4 Combination of Machine Scores

To build an accurate and robust predictor of human ratings, several different machine scores may need to be combined. The performance of a few methods of combination were tested by Franco et al. [10], while Cincarek et al. [11] used combinations of scores to classify words as correctly pronounced or mispronounced. We describe the work by Franco et al. in more detail here.

### 2.4.1 Methods of Combination

Franco et al. present four different methods by means of which machine scores can be combined. The rating a human would assign to an utterance is viewed as a random variable, and the goal is to estimate or predict the value of this human rating,  $h$ , using a set of machine scores as predictors.

#### Linear Combination

This approach assumes that the ideal human rating can be approximated as a linear combination of machine scores:

$$h = a_1m_1 + a_2m_2 + \dots + a_nm_n + b,$$

where  $m_1, m_2, \dots, m_n$  represent  $n$  different machine scores.

The linear coefficients  $a_1, a_2, \dots, a_n, b$  are chosen by means of linear regression to minimise the mean square error between the predicted rating and the actual human rating, based on a set of training data. This is a reasonably simple approach and leads to robust estimates. (Franco et al. [10]).

#### Artificial Neural Networks

Neural networks are a promising method of combination if the relationship between machine scores and human ratings are severely non-linear. The different machine scores form the input of a neural network that computes the non-linear mapping  $o(\ )$  of these scores to a predicted human rating  $h$ :

$$h = o(m_1, m_2, \dots, m_n)$$

The neural network can be trained iteratively, aiming to minimise the mean square error between the predicted and actual human ratings. However, there is a risk of overfitting the

network to the specific training data, making it less robust. To counter this effect a second data set, the validation set, is used. Training is done based on the training set, and halted when performance ceases to improve on the validation set.

Neural networks are difficult to interpret, and are computationally expensive to train. Franco et al. report having to make a large number of manual adjustments in order to create an effective neural network for combining machine scores [10].

### Distribution Estimation

In this method the expected human rating is calculated using estimates of the conditional probabilities  $P(h_i|m_1, \dots, m_n)$ . The expected human rating is then

$$h = \sum_{i=1}^G h_i \cdot P(h_i|m_1, m_2, \dots, m_n),$$

where  $G$  is the number of distinct discrete human ratings that could be assigned. Using Bayes' Rule, we can express the above conditional probability as

$$P(h_i|m_1, m_2, \dots, m_n) = \frac{P(m_1, m_2, \dots, m_n|h_i)P(h_i)}{\sum_{j=1}^G P(m_1, m_2, \dots, m_n|h_j)P(h_j)}.$$

The densities  $P(m_1, m_2, \dots, m_n|h_i)$  are approximated by discrete distributions which in turn are estimated using a quantisation of the machine scores. Scalar or vector quantisation can be used. For scalar quantisation, Franco et al. [10] experimented with using different numbers of bins on a set of training data, calculating the correlation with human ratings in each case. This allowed the authors to select the optimal number of bins for combining three different machine scores, two different machine scores or using a single machine score. It was found that the correlation with human ratings fell when too many or too few bins were employed. For the vector quantisation case, Franco et al. [10] again experimented with different numbers of codewords, finding the optimal number of codewords for combining three scores, two scores or just modeling a single score. Codewords were designed using the Linde-Buzo-Gray algorithm.

Although distribution estimation using vector quantisation was found to be one of the more successful methods of combination, the authors note that much experimentation was required to set up an effective system.

### Regression Trees

A second approach to the estimation of the probability  $p(h|m_1, m_2, \dots, m_n)$  is using classification and regression trees. Such a tree takes a vector of machine scores as input. Starting at the root of the tree, a child-node is chosen at each parent node based on the machine score vector, until a leaf node is reached. Each leaf node corresponds to a different human rating.

Franco et al. generated trees using a public domain software package, minimising the mean square error computed over a set of training data. The authors note that, compared to neural networks and distribution estimation, trees are quick and simple to create and interpret [10].

### 2.4.2 Results

Franco et al. [10] used three different machine scores described in the study by Neumeyer et al. [7] for experimenting with combination methods. These scores were the *posterior HMM log-likelihood*, the *segment duration score* and the *syllabic timing score*. The speaker level correlations of these scores with human ratings are shown in Table 2.1. However, for the combination experiment, Franco et al. used sentence level correlations. Of these three raw scores, the posterior HMM-LL had the highest sentence level correlation with human ratings, 0.58. This was used as a baseline for evaluating the performance of the different combination techniques. The correlations of the *segment duration score* and the *syllabic timing score* with human ratings were 0.47 and 0.35 respectively. The three different machine scores had correlations of between 0.43 and 0.66 with each other, implying that they each contain some amount of independent information.

For each combination method, the non-native speech data was divided into two equally sized sets, one used for training the parameters of the combination method, and the other used for testing the method’s performance. The correlation between the ratings produced by combination and the assigned human ratings was then calculated. Finally, the training and testing sets were swapped, the process repeated, and the average of the two correlations taken.

Linear combination of the HMM-LL and segment duration scores showed a slight increase in performance over the baseline. Adding syllabic timing as a third input led to another slight improvement. However, none of these performance increases were of significant magnitude.

Combination using a neural network was most successful, with the optimal configuration showing a correlation of 0.64, an increase of 11.5% over the baseline (posterior HMM-LL) correlation. Using the neural network to create a non-linear mapping of the posterior HMM-LL alone increased the correlation with human ratings by 8%. Combining the posterior HMM-LL and the segment duration scores led to a 10.8% improvement over the baseline, while the combination of posterior HMM-LL, segment duration scores and syllabic timing resulted in the full improvement of 11.5% over the linear use of posterior HMM-LL alone.

Distribution estimation using scalar quantisation improved the correlation with human scores by 6.1% above the baseline when only using posterior HMM-LL. The addition of the other two scores resulted in decreased correlation with human ratings.

Distribution estimation using vector quantisation was more successful, providing an increase in correlation of 7.3%. Overwhelmingly, this increase is due to the non-linear mapping of the posterior HMM-LL. The addition of the two other scores resulted in only a marginal increase in correlation.

Finally, regression trees resulted in an increase in correlation of 8% when combining all three scores.

The results show that the non-linear mapping of a single strongly correlated machine score provides a substantial improvement in correlation with human ratings. It is possible to increase this correlation somewhat by combining more machine scores. For this purpose neural networks show the most promise. Regression trees are a simpler alternative that still results in significant improvement of correlation with human ratings. [10]

The performance of the above methods of combination are summarised in Table 2.2.

Score Name	Combinations		
<i>Posterior HMM-LL</i>	✓	✓	✓
<i>Segment Duration Score</i>		✓	✓
<i>Syllabic Timing Score</i>			✓
Method	Improvement		
Linear Combination	<i>Baseline</i>	1.9%	3.0%
Neural Networks	8.0%	10.8%	11.5%
Distribution Est. (Scalar)	6.1%	5.0%	-1.4%
Distribution Est. (Vector)	6.8%	7.1%	7.3%
Regression Trees	5.7%	7.3%	8.0%

**Table 2.2:** Performance of machine score combination methods relative to the correlation of posterior HMM-LL with human ratings.

## 2.5 Summary and Conclusions

This chapter has given an overview of previous research in the field of automated oral proficiency assessment. We described a number of studies that were most relevant to the proposed research and the machine scoring algorithms they employ. We also introduced different methods of combining machine scores.

Studies vary in their assessment strategies and data composition. This complicates the comparison of machine score performance between them. The selection of scoring algorithms for the research described in this thesis is further complicated by the proficiency level of the intended test population. While the studies described in this chapter investigate the assessment of *foreign language* speakers, the focus of our research is the assessment of *second language* speakers. For foreign language speakers, the use of the target (L2) language can be seen as limited to the classroom, while second language speakers use the L2 language in their daily lives [5]. Therefore, second language speakers can be expected to be more proficient in the L2 language than foreign language speakers, and scoring algorithms which

appear promising based on an experiment involving one group may not be equally effective in an experiment involving the other.

Eleven machine scoring algorithms were selected for this research. In Chapter 5 we investigate Witt & Young's *Goodness of Pronunciation* score as an established scoring algorithm based on HMM log-likelihood. In Chapter 6 we employ *Rate of Speech*, which has been shown to be a simple and robust measure of oral proficiency, as well as the related scores *Articulation Rate* and *Phonation/Time Ratio*. We also examine the *Segment Duration Score*, which resulted in strong correlations with human ratings in the studies by Neumeyer et al. [7] as well as Hacker et al. [9]. Lastly, three scores based on the accuracy of a repeated utterance are investigated in Chapter 7.

Finally, for assessing combinations of these scores, we choose linear regression due to its simple and intuitive implementation. This is described in Chapter 8.



# Chapter 3

## Data Corpus

To evaluate the potential of different machine scores to predict human assessments of oral proficiency, we require a corpus of speech that has been evaluated by human raters.

At the Stellenbosch University Faculty of Education, students enrolled for the “Postgraduate Certificate in Education” require a language endorsement [4]. Many of these students are Afrikaans mother tongue speakers with English as a second language. They must enrol for a language module appropriate to their level of proficiency, and their progress must be monitored regularly. These students were used as the test population for this study.

The students took an automated oral test and some of their answers were rated for oral proficiency by a group of human raters. This chapter presents the design of the test, the rating process, and the rating scales used to evaluate responses. We investigate the quality of the corpus by determining inter-rater agreements, intra-rater correlations and inter-scale correlations. Finally, we describe the method used to compare machine scores to the human proficiency ratings assigned for this corpus.

### 3.1 Test Description

A computerised test was used to collect responses from students at the Faculty of Education. The aim of the test was to assess listening and speaking skills in the context of secondary school education. Therefore, the contents of the test relates to language use in this domain and no attempt was made to mimic natural human dialogue [4].

The test was implemented over the telephone. This method requires a minimum of specialised equipment and allows the test to be taken from any number of different locations. For this experiment, calls were placed from a telephone located in a private office reserved for this purpose.

Students were guided through the test by a spoken dialogue system. This system did not interpret replies by students, but merely played prompts based on a pre-defined test structure and recorded students’ answers for later, off-line processing.

The system’s spoken prompts were recorded using different voices for test guidelines, instructions and examples of proper responses, to make the test easy to follow. Students

received both oral instructions before the test and a printed test sheet with instructions and certain prompts.

### 3.1.1 Test Design

The complete test consisted of seven tasks, each requiring the student to comprehend the instructions spoken by the system and to respond verbally. In this research we focus on only two of these tasks, namely the *reading task* and the *repeating task*. For a description of the complete test, the reader is referred to [1].

#### Reading Task

Students received a printed test sheet with eleven sentences to be read for the reading task. For each student, six of these sentences were selected at random by the system. The student was prompted to read each in turn, and the resulting utterances were recorded. As an introduction, the system played an example response before prompting for the first sentence. The sentences used for the reading task are listed in Appendix A.

This task was familiar to students, since it is similar to parts of their secondary school language examinations. Relying on the printed test sheet was intended to help nervous candidates to relax [3].

#### Repeating Task

In this task students were instructed to listen to a prompt played by the system, and then repeat the same sentence. As before, the system played an example prompt and response before starting the repeating task. The eight sentences used for this task are listed in Appendix B. Students were prompted to repeat each of these sentences in random order.

The task design is based on the hypothesis that oral production is influenced by the student's phonological working memory. The expectation is that during oral communication, second language speakers would struggle to produce the desired utterance due to their limited access to the vocabulary and sound system of the target language (see [4] and references therein).

### 3.1.2 Test Population

The test was taken by 120 students as part of their oral proficiency assessment. The majority of these students are Afrikaans mother tongue speakers, whose proficiency in English varies from intermediate to advanced. Feedback from the students indicated that most of the Afrikaans-speaking students found the test challenging, while the English-speaking students found it manageable [3].

Of the 120 students, 90 were selected to form a test set, which was representative of the gender and first language composition of students at the Faculty of Education. The results

in this study are based on this set of 90 students. Of the remaining students, 16 were selected to form a development set, which was used to tune the recogniser and certain machine score algorithms.

## 3.2 Human Ratings

Human perceptions of the test population’s oral proficiency are central to this research. When developing machine scoring algorithms, we aim to predict with reasonable accuracy the proficiency ratings assigned to the recorded utterances by human raters. Furthermore, the agreement among the different raters and their individual rating consistencies serve as a benchmark against which we can compare the performance of an automatic scoring system.

In initial experiments conducted with this corpus, raters assigned each student a single proficiency rating for the reading task and a single rating for the repeating task. These ratings were based on two separate five-point Likert scales, one for each task [3].

However, for the experiments described in this research, the scales were redesigned, resulting in an improvement over the initial experiments in terms of rater consistency and agreement. The revised scales separate certain aspects of proficiency and are more detailed than those used in the initial experiment. This research relies only on the proficiency ratings obtained using this refined set of scales. For a detailed discussion of the initial experiments, see [3].

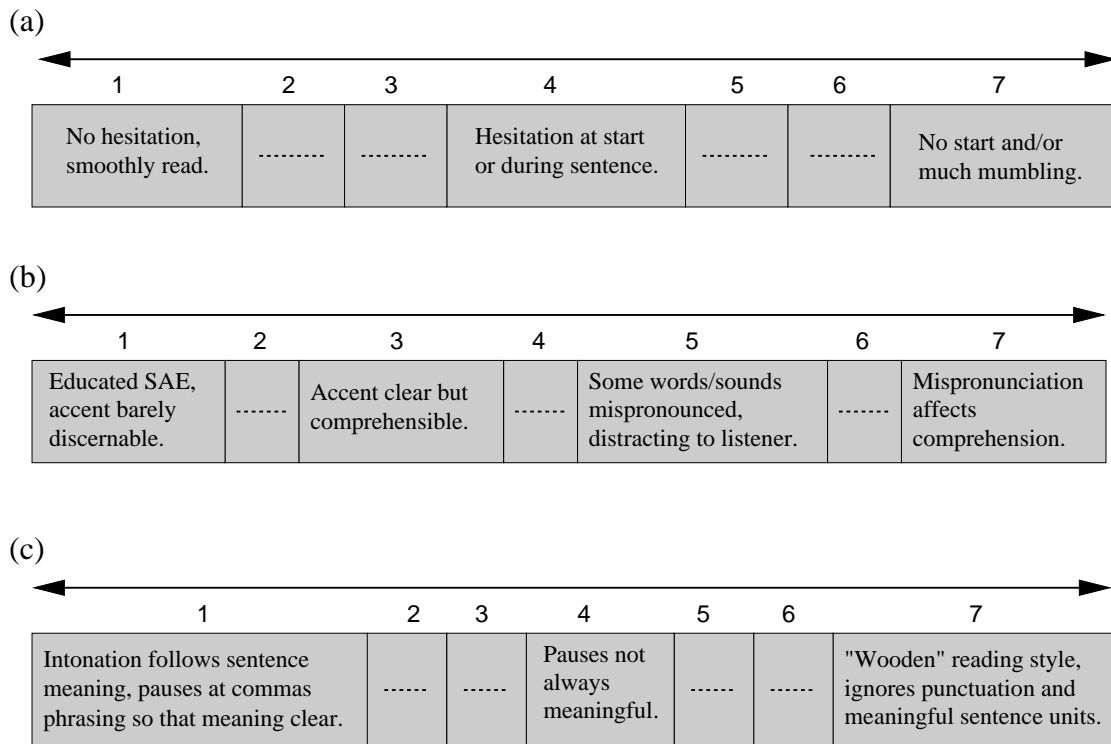
### 3.2.1 Rating Scales

Five different scales were designed, each aimed at evaluating a different aspect of oral proficiency. *Hesitation*, *Pronunciation* and *Intonation* were used to assess the reading task. The corresponding scales are shown in Figure 3.1. *Success* and *Accuracy* were used to evaluate the repeating task, and the scales are shown in Figure 3.2. Raters were required to assign multiple ratings to each utterance, one from each of the relevant task’s rating scales.

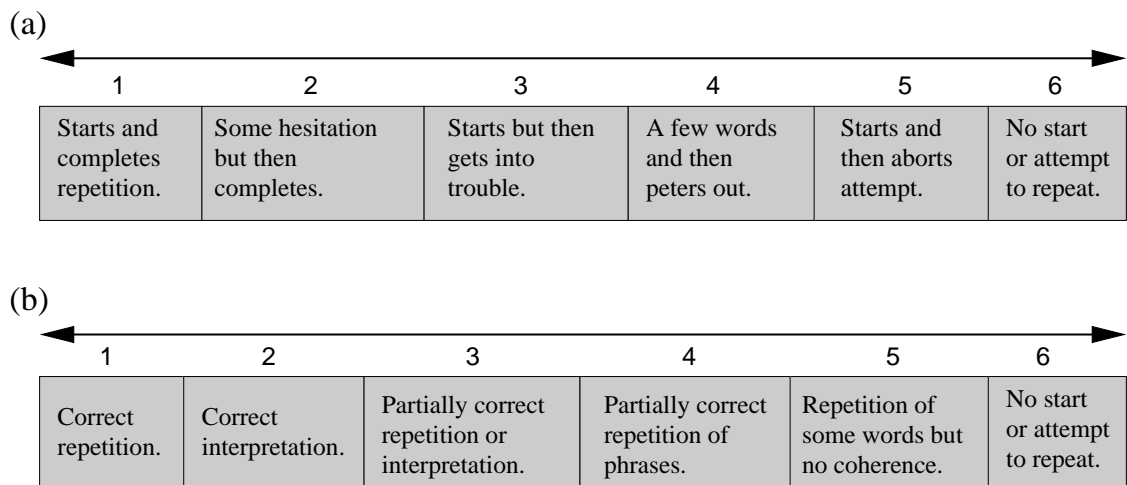
Feedback from the initial experiments had indicated that raters sometimes experienced the Likert scales as too restrictive and wanted to assign a rating between two adjacent Likert points. Therefore, the new reading task scales included some unlabelled Likert points. The numbers above the scales in Figures 3.1 and 3.2 show the rating values associated with each point. These numbers were not included on the scales given to the raters, to avoid prior perceptions of “good” or “bad” marks from influencing the ratings.

### 3.2.2 Human Raters

Six teachers of English as a second or foreign language were asked to rate the student responses recorded during the test using the scales described above. The raters did not know the students personally, and each had approximately the same level of training and



**Figure 3.1:** Reading task rating scales used to assess (a) degree of Hesitation, (b) Pronunciation and (c) Intonation. Adapted from [4].



**Figure 3.2:** Repeating task rating scales used to assess (a) degree of Success and (b) Accuracy. Adapted from [4].

experience. A short initial training session was offered, where some example utterances were played and the use of each scale was discussed [3].

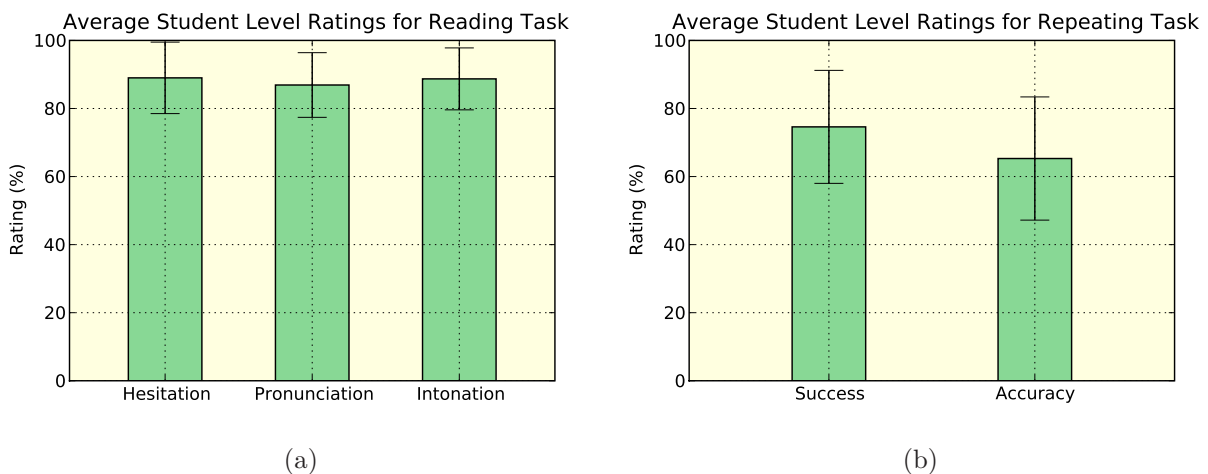
Each student’s responses were assessed by three different raters. This allowed the *inter-rater agreement* to be calculated, which indicates the extent to which the raters agreed about the ratings assigned to each utterance.

Each rater assessed 45 different students, five of whom were presented to the rater twice. This allowed the *intra-rater correlation* to be calculated, as a measure of the rater’s consistency in assigning the same ratings to the same utterance.

Due to limited manpower and resources, it was not feasible to rate all the student responses. Instead, three reading task responses and three repeating task responses were chosen at random for each of the 90 students in the test set.

### Average Ratings

Figures 3.3(a) and 3.3(b) show the mean ratings assigned for each scale of the reading and repeating task respectively. The standard deviation is indicated in each case by the horizontal bars. In the figures, ratings are presented as percentages, to simplify interpretation and comparison.



**Figure 3.3:** Mean ratings assigned for each of the (a) reading and (b) repeating task rating scales. Horizontal bars show the standard deviations.

The high mean ratings for all three of the reading task scales seem to indicate that students did not find the reading task sufficiently challenging. This is supported by the low standard deviations, showing that ratings for the reading task were concentrated in the upper region of the rating scales. It is likely that a future iteration of the test would benefit from more challenging reading task prompts.

The ratings for both scales of the repeating task have lower means and higher standard deviations than those of the reading task. This leads us to conclude that the repeating task was of the appropriate difficulty level for this test population.

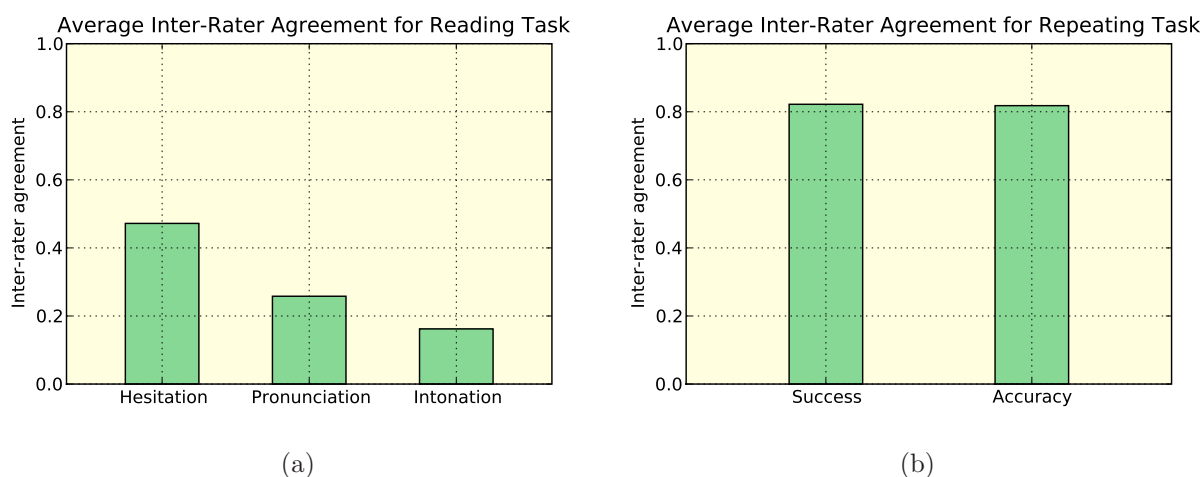
### Rater Consistency and Inter-Rater Agreement

Each rater's intra-rater correlation was calculated based on the ratings assigned to the five students who were evaluated twice. The correlations are two-way random, intra-class correlation coefficients and were calculated using Statistica [14]. The correlations were based on both the reading and repeating task ratings, and are shown in Table 3.1. The average of these correlations, 0.85, compare favourably with those reported in other studies [10; 8].

Rater	Intra-rater correlation
1	0.83
2	0.94
3	0.81
4	0.96
5	0.67
6	0.91

**Table 3.1:** *Intra-rater correlations for human raters. Adapted from [4].*

Figures 3.4(a) and 3.4(b) show the average inter-rater agreement for each of the rating scales. These values reveal how well the raters agreed on the ratings assigned for each utterance.



**Figure 3.4:** *Average inter-rater agreement for each of the (a) reading and (b) repeating task rating scales.*

It is clear from Figure 3.4(a) that raters differed to a large extent in their assessments of the reading task. While this may be due in part to the design and definition of the relevant rating scales, it is also believed to be related to the high means and low standard deviations for the reading task ratings, shown in Figure 3.3(a). Because the students generally performed very well in this task, there is little contrast between high and low proficiency re-

sponses, making it difficult for raters to be consistent in their assessments of these utterances. A similar observation was made by Zechner et al. [15].

It is important to note that this same phenomenon will be found to play a role in causing relatively low correlations between machine scores and human ratings for the reading task throughout this study.

In contrast to the reading task, the average ratings assigned for the repeating task scales have lower means and higher standard deviations, as shown in Figure 3.3(b). The fact that students struggled more with the repeating task and received ratings that are better spread throughout the scales than those for the reading task, makes it more likely that raters will agree on the ratings assigned to each utterance. This can be seen in the high average inter-rater agreements for the repeating task, shown in Figure 3.4(b).

### Inter-Scale Correlations

We estimate the importance of each of the five rating scales by considering their correlations with each other. These correlations are shown in Table 3.2.

The table also shows the correlations of each rating scale with the students' academic oral and progress marks for the relevant course during the academic year in which the test was taken. The oral mark is based on the lecturer's assessment of a number of oral exercises performed during the course, including prepared presentations and role-playing situations. The progress mark is composed of assessments for written work, the oral mark, as well as a written test.

	Oral Marks	Progress Marks	<i>Hesitation</i> Ratings	<i>Pronunciation</i> Ratings	<i>Intonation</i> Ratings	<i>Success</i> Ratings	<i>Accuracy</i> Ratings
Oral Marks	1.00	0.49	<b>0.11</b>	0.34	<b>0.03</b>	0.41	0.43
Progress Marks		1.00	0.16	0.27	0.09	0.27	0.25
<i>Hesitation</i> Ratings			1.00	0.40	<b>0.66</b>	0.24	0.26
<i>Pronunciation</i> Ratings				1.00	0.41	0.34	0.38
<i>Intonation</i> Ratings					1.00	0.27	0.24
<i>Success</i> Ratings						1.00	<b>0.89</b>
<i>Accuracy</i> Ratings							1.00

**Table 3.2:** *Correlations between rating scales and students' academic marks.*

The table reveals that ratings for *Hesitation* and *Intonation* have almost no correlation with the students' oral marks. These two rating scales also have a relatively high correlation with each other. Our focus with regard to the reading task will therefore be on the *Pronunciation* ratings.

The ratings for the two repeating task scales, *Success* and *Accuracy*, are highly correlated. In a future experiment, it may be sufficient to use only one of these two scales.

Overall, the correlations between the oral proficiency ratings and the oral academic marks are disappointingly low. It is possible that these correlations may be improved by using more data per student than the three utterances from each task that have been considered here. However, these low correlations can also be attributed to the fact that the tasks performed to obtain the oral academic marks are different from the test used to obtain the oral proficiency ratings. This leads us to conclude that the proficiency ratings outlined here and the academic oral mark determined during the academic course evaluate different aspects of students' ability to express themselves verbally.

### 3.3 Evaluation of Machine Scores

Throughout this thesis, the term “*scores*” refers to automatically derived machine scores, while “*ratings*” refers to human assessments.

The machine score algorithms presented in this study were evaluated by determining how strongly they are correlated with each of the human rating scales. The aim is to identify which scores have the potential to predict which rating scales.

#### 3.3.1 Averaging Human Ratings

The correlation between machine scores and human ratings can be calculated at the *utterance level* or at the *student level*.

Each utterance was assessed by three of the six human raters. The ratings by these three raters were averaged to give the utterance level ratings for each of the task's scales. Where a student's responses were presented twice to the same rater, the mean rating assigned by that rater was determined first. *Utterance level correlations* can then be calculated by comparing these ratings with the machine scores calculated for each utterance.

For each student, three responses to each task were assessed. By averaging the utterance level ratings for these three responses, we can calculate the student level ratings for each scale. In the same manner, student level machine scores were calculated by determining the average scores assigned to each utterance. The *student level correlations* were calculated by comparing the average ratings with the average machine scores for each student.

Preliminary investigations revealed that student level correlations are generally higher than utterance level correlations when used to evaluate machine scores. This seems to indicate that the agreement between machine scores and human ratings improves when



more information about each student is available. Therefore, we consider only student level correlations in this study.

### 3.3.2 Spearman’s Rank Correlation Coefficient

We expect the machine scores calculated in this study to be ordinal, but not necessarily normally distributed. We therefore use Spearman’s rank correlation coefficient to calculate the correlation between machine scores and human ratings.

To calculate the Spearman’s rank correlation coefficient of two data vectors, each vector is first ranked numerically. If two or more data points have the same rank, they are assigned the average of their positions, as shown in the example below. Spearman’s rank correlation coefficient is then determined by calculating Pearson’s correlation coefficient for the ranked data [16].

For two variables  $x$  and  $y$  with  $n$  instances, Pearson’s correlation coefficient,  $r$ , is defined as:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Table 3.3 shows an example of the ranks assigned to the machine scores and average ratings of a hypothetical set of 6 students. Students 1 and 6 received the same machine score, 0.4. They therefore share the first and second positions, and are assigned the average rank of those positions, 1.5. Similarly, students 3, 4 and 6 received the same human rating of 4. They are therefore tied for positions 2, 3 and 4 and are assigned the average of those ranks, 3. Calculating Pearson’s correlation coefficient using these rank values results in a Spearman’s coefficient of 0.59. The unranked Pearson’s correlation coefficient value is 0.50.

Student	Machine Score		Human Rating	
	Value	Rank	Value	Rank
1	0.4	<b>1.5</b>	2	<b>1</b>
2	1.6	<b>4</b>	6	<b>5</b>
3	3.1	<b>6</b>	4	<b>3</b>
4	0.7	<b>3</b>	4	<b>3</b>
5	2.2	<b>5</b>	7	<b>6</b>
6	0.4	<b>1.5</b>	4	<b>3</b>

**Table 3.3:** Example of ranks for calculating Spearman’s rank correlation coefficient.

Because Spearman’s correlation requires the data points to be ranked before determining Pearson’s coefficient, the effects of non-linearities in the data are greatly reduced. This allows

us to better evaluate the potential of machine scores. If these scores are in fact non-linearly related to human ratings, a completed automatic assessment system could apply non-linear transformations to the scores or use non-linear methods of combining machine scores to predict human ratings.

### 3.4 Summary and Conclusions

This chapter has described the corpus of data that was used to conduct the experiments described in this thesis. This corpus consists of a set of recorded oral exercises performed by a set of 90 students. These exercises were subsequently rated for oral proficiency by six independent human raters according to a well-defined set of scales.

We evaluated the reliability of the human raters by calculating the inter- and intra-rater correlations. Table 3.1 shows that raters were reasonably consistent in their ratings. The high average ratings for the reading task, shown in Figure 3.3(a), contributed to low agreement among raters for that task, shown in Figure 3.4(a). Inter-rater agreement for the repeating task scales were higher, as shown in Figure 3.4(b)

Finally, we described the calculation of student level ratings and scores, as well as the use of Spearman's rank correlation coefficient for evaluating the potential of machines scores to predict human ratings of oral proficiency.

# Chapter 4

## Automatic Speech Recognition System

The machine scoring algorithms investigated in this study all rely on features that can be extracted from test utterances by means of automatic speech recognition (ASR). Some require the automatic transcription, some the recognition probabilities of individual phones, and others the phonetic segmentation of the utterances to be scored.

In this chapter we describe the automatic speech recogniser used and discuss the five different recognition strategies on which this study's results are based.

### 4.1 Recogniser

The Hidden Markov Model Toolkit (HTK) version 3.4 was used for ASR in this project [17]. This toolkit was developed and is licensed and maintained by the Cambridge University Engineering Department. It is freely available online from [htk.eng.cam.ac.uk](http://htk.eng.cam.ac.uk).

The hidden Markov models (HMMs) used by the speech recogniser were trained on approximately 6 hours of telephone quality speech. This data is part of the African Speech Technology (AST) corpus, and consists of phonetically and orthographically annotated speech gathered over South African fixed as well as mobile telephone networks [18]. Triphone HMMs were obtained by means of decision-tree state clustering and embedded Baum-Welsh re-estimation. The final set of triphone HMMs consisted of 4797 tied states based on a set of 52 phones, and a maximum of 8 Gaussian mixtures per HMM state.

The HMM training did not form part of this work, since the models were available from previous projects within the Department.

#### 4.1.1 Recognition Output

Recognition was performed using the HTK tool *HVite*, which performs Viterbi decoding based on a set of HMMs and a language model, grammar or reference transcription, depending on whether a forced alignment is being performed or not. *HVite* produces label files

which contain the ASR transcription, the start and end times for each phone, as well as the average log probability per frame for each phone.

From the latter, the average log probability per frame for phone  $q_i$ , which spans  $N_i$  frames, is calculated as

$$p_{q_i} = \frac{\log(p(O_i|q_i))}{N_i}$$

where  $p(O_i|q_i)$  is the probability of observing the  $i_{th}$  acoustic segment  $O_i$  given the phone model  $q_i$ . This quantity is defined at the frame level for  $N_i$  frames as:

$$p(O_i|q_i) = \prod_{n=1}^{N_i} p(s_{i_n}|s_{i_{n-1}})p(o_{i_n}|s_{i_n}) \quad (4.1)$$

where  $s_{i_n}$  denotes the state of the HMM for phone  $q_i$  associated with  $o_{i_n}$ , the  $n_{th}$  observation in acoustic segment  $O_i$ , and  $p(s_{i_n}|s_{i_{n-1}})$  denotes the HMM transition probability between states  $s_{i_n}$  and  $s_{i_{n-1}}$ . The value of  $p(O_i|q_i)$  is calculated as part of the Viterbi decoding process by the HTK tools.

## 4.2 Recognition Strategies

Five different speech recognition strategies were used during the course of this project:

1. Finite state grammar.
2. Unigram language model.
3. Oracle finite state grammar.
4. Oracle alignment.
5. Free phone loop grammar.

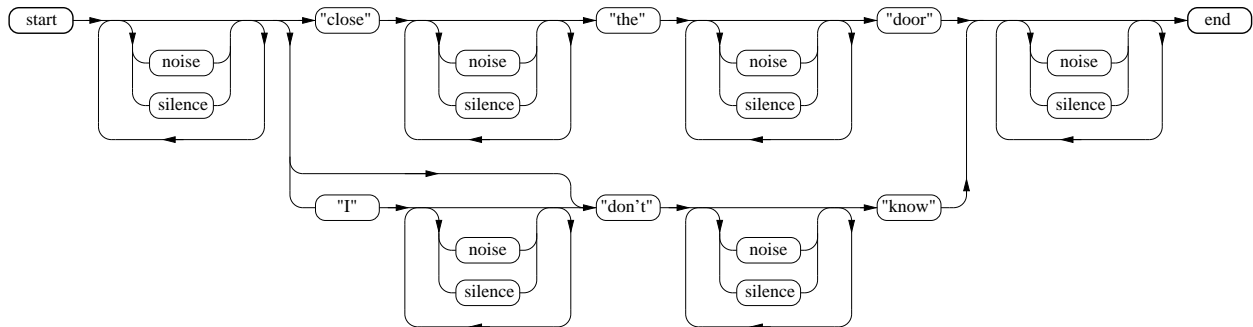
These strategies differ in the grammar used during decoding. The first two can be derived automatically, and the focus will be on the results based on these two approaches when evaluating the performance of machine scoring algorithms. The second two are based on human transcriptions. They are therefore not automatically realisable, but give an indication of the impact better recognition accuracies would have on the various machine scores. The last approach is required by the posterior log-likelihood scoring algorithms described in Chapter 5.

### 4.2.1 Finite State Grammar

This grammar was used for the automatic recognition of the *reading task*. It is expected that the students, who generally have good English reading skills, would make very few

word errors while reading prompts from a test sheet. Hence the use of a strict finite state grammar (FSG) is an appropriate recognition method for this task.

For each prompt of the reading task an FSG was created allowing the desired utterance, “I don’t know”, or simply “don’t know”. The branch allowing the desired utterance expects all words to be present. Silence, noise and hesitation sounds are allowed between words. Figure 4.1 shows an example of such a network.



**Figure 4.1:** Example of a finite state grammar network for the hypothetical sentence “Close the door.”

These prompt-specific grammars were defined using extended Backus-Naur form (EBNF) notation and parsed using the HTK tool *HParse* to form lattice files, which were then used by *HVite* during the recognition process. The process is described in detail in [17].

### 4.2.2 Unigram Language Model

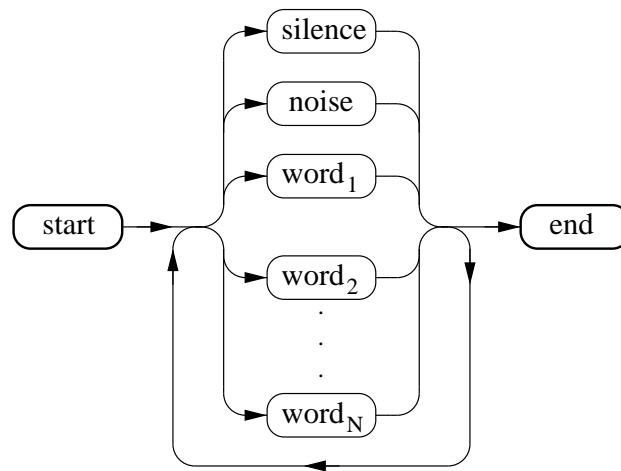
This recognition strategy was used for automatic recognition of the *repeating task*. For this task, provision must be made for missing words, changes in word order, and the replacement of words or phrases with synonyms. This makes the use of a strict FSG less attractive than the use of a unigram language model (LM), which places no restrictions on word order.

The unigram LM consists of a word loop, where the allowed words are obtained from the human transcriptions of the development set as well as all words occurring in the respective prompt. Silence and noise are allowed between words. Figure 4.2 illustrates this structure. A separate LM was created for each prompt of the repeating task.

The word loop was unweighted, meaning that all word-to-word transitions had equal probability.

### 4.2.3 Oracle Finite State Grammar

Recognition using an oracle FSG is similar to recognition using an FSG grammar as described above. However, instead of creating a grammar for each prompt based on the desired utterance, grammars are created for each utterance based on the human transcription of the actual utterance. As before, silence, noise and hesitation sounds are allowed between words.



**Figure 4.2:** Network showing the structure of the unigram LM recognition strategy.

While human transcriptions can be seen as the most accurate representation of the words in an utterance, non-speech events such as silence and noise are often not accurately transcribed. The use of an oracle FSG is aimed at obtaining recognition output with a zero word-error rate, but also including silence and noise where appropriate.

This recognition strategy was used for speech recognition in both the reading and repeating tasks. Results based on this method are used as an indication of the effects that improved automatic speech recognition would have on the various machine scoring algorithms presented.

#### 4.2.4 Oracle Alignment

The term *oracle recogniser* refers to a hypothetical ideal speech recogniser. For the reading task, the responses of 10 students were meticulously transcribed by hand to form a set of oracle recognition transcriptions. In these transcriptions, care was taken to transcribe words as well as non-speech events accurately.

The students in question were the 5 with the highest and the 5 with the lowest human ratings for *Pronunciation*. Students were selected in this way to create a set with the maximum contrast between high and low proficiency. Machine score performance based on the oracle transcriptions of these students' responses can be seen as a best-case-scenario, where students vary greatly in proficiency and automatic recognition is ideal.

#### 4.2.5 Free Phone Loop Grammar

Free phone loop recognition is based on a simple FSG that allows an arbitrary sequence of phones and silences, with no context restrictions or prior probabilities. It was performed for both the reading and repeating tasks during the calculation of the posterior log-likelihood algorithms of Chapter 5.

# Chapter 5

## Posterior Log-Likelihood Scoring

In this chapter we investigate the correlations between human ratings for our test data and a variety of posterior log-likelihood scores.

Likelihood scores focus on the *acoustic* characteristics of the utterance to be scored, rather than on temporal, or segmentation-related, characteristics. Scores are calculated using the recognition likelihood for each recognised phone. Neumeyer et al. argue that such likelihood scores can be adversely affected by spectral mismatch between the recogniser models and the utterance under investigation, due to speaker and acoustic channel characteristics that are unrelated to the speaker’s oral proficiency [7]. Posterior log-likelihood scoring is proposed as a more robust scoring measure, and is expected to be less affected by such spectral mismatch. The posterior log-likelihood is calculated as a ratio between the recognition likelihood of a phone selected by forced alignment and the recognition likelihood of a phone selected by free phone loop recognition. It is argued that, since the effects of any spectral mismatch would be present in both likelihoods, the score would be less affected by this mismatch.

We will refer to our posterior log-likelihood score as the *Goodness of Pronunciation* (*GOP*) score, after Witt & Young [6]. While Witt & Young used *GOP* scores to accept or reject individual phones based on pronunciation quality, we will calculate utterance level scores for comparison with the human ratings of the test sentences.

We first discuss the algorithm used to calculate *GOP* scores, then we identify four methods of combining phone level *GOP* scores to form utterance level scores, and finally we present the correlations of these scores with the various human rating scales applied to our data.

## 5.1 Score Definitions

We base our phone level *GOP* score algorithm on Equation 2.1 as defined in [6]. The score can be expressed as follows:

$$\begin{aligned}
 GOP(q_i) &= \left| \log \left( \frac{p(O_i|q_i)}{\max_{j=1}^J p(O_i|q_j)} \right) \right| / N_i \\
 &= \left| \frac{\log(p(O_i|q_i))}{N_i} - \frac{\log(\max_{j=1}^J p(O_i|q_j))}{N_i} \right| \\
 &= |p_{q_i(\text{forced})} - p_{q_i(\text{free})}|
 \end{aligned} \tag{5.1}$$

where  $N_i$  is the number of frames composing acoustic segment  $O_i$ , and  $p(O_i|q_i)$  is the probability of observing  $O_i$  given the phone model  $q_i$ , as defined in Equation 4.1.

The terms in Equation 5.1 can be extracted from the recognition output of the HTK *HVite* function. The term  $\frac{\log(p(O_i|q_i))}{N_i}$  corresponds to the average log probability per frame for the phone  $q_i$ , obtained from a forced alignment,  $p_{q_i(\text{forced})}$ , between the acoustic features and the expected transcription of the utterance. The term  $\frac{\log(\max_{j=1}^J p(O_i|q_j))}{N_i}$  corresponds to the average log probability for the same frames, calculated using free phone loop recognition,  $p_{q_i(\text{free})}$ .

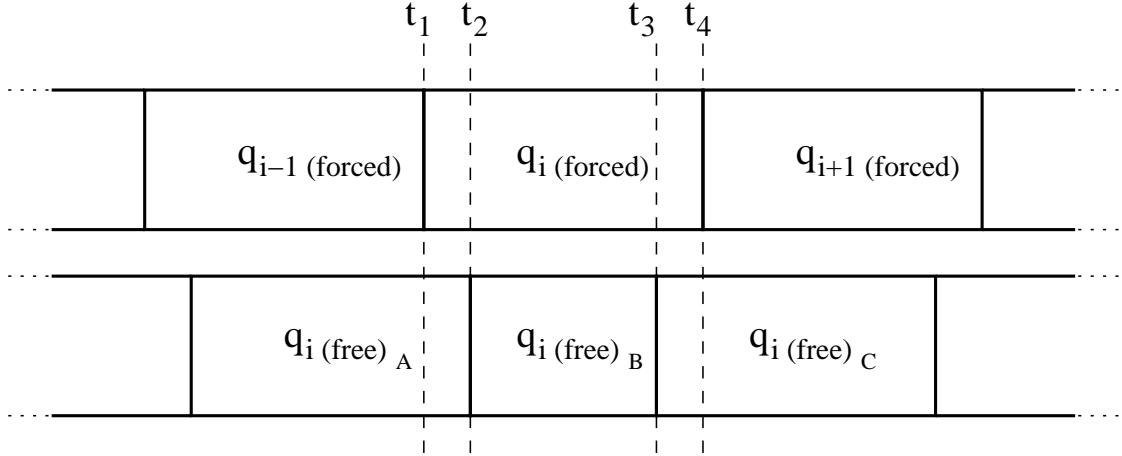
During a forced alignment, the recogniser matches acoustic segments to the phones determined by a reference transcription or finite state grammar. Free phone loop recognition allows the recogniser to match phones to acoustic segments without any grammatical restrictions. The recognition strategies are described in Section 4.2.

From Equation 5.1 we see that a poorly pronounced phone  $q_i$  would lead to a large difference between the forced and the free-phone scores, leading to a high *GOP* score. For well-pronounced phones, a low score is expected.

The utterances to be scored are therefore recognised twice, once using a free phone loop and once using forced alignment. The *GOP* score for each phone in the forced alignment is calculated in turn. For each of these phones, those selected by a free phone loop recognition and which span the same frames, are identified. The segmentation in the free phone loop recognition will in general differ from the segmentation of the forced alignment. Hence the free phone loop average log probabilities are weighted by the duration of the overlapping part of the segment, before being subtracted from the average log probability of the force aligned phone segment. An example of this process is shown in Figure 5.1. Equation 5.2 indicates how the *GOP* for phone  $q_i$  is calculated.

$$\begin{aligned}
 GOP(q_i) &= |p_{q_i(\text{forced})} - p_{q_i(\text{free})}| \\
 &= \left| p_{q_i(\text{forced})} - \left( \frac{t_2 - t_1}{t_4 - t_1} p_{q_i(\text{free})_A} + \frac{t_3 - t_2}{t_4 - t_1} p_{q_i(\text{free})_B} + \frac{t_4 - t_3}{t_4 - t_1} p_{q_i(\text{free})_C} \right) \right|
 \end{aligned} \tag{5.2}$$





**Figure 5.1:** *Mismatched segmentation between phones selected by forced alignment,  $q_i \text{ (forced)}$ , and those selected by free phone loop recognition,  $q_i \text{ (free)}$ . Three phones selected by free phone loop recognition,  $q_i \text{ (free)}_A$ ,  $q_i \text{ (free)}_B$  and  $q_i \text{ (free)}_C$ , overlap with  $q_i \text{ (forced)}$ , the  $i^{\text{th}}$  phone selected by forced alignment.*

In order to compare the  $GOP$  scores with human ratings, utterance level scores must be obtained from the phone level  $GOP$  scores. These utterance level scores are calculated by totalling the phone level  $GOP$  scores of the force aligned phones in the utterance and normalising by the number of phones,  $M$ :

$$GOP = \frac{\sum_{i=1}^M GOP(q_i)}{M}$$

Several variations of the utterance level  $GOP$  score can be obtained by specifying the types of phones included in the above calculation.

### 5.1.1 All Phones - $GOP_{All}$

In its simplest form, the utterance level  $GOP$  score can be calculated by using all force aligned phones [7]. This  $GOP$  score will be referred to as  $GOP_{All}$ .

### 5.1.2 Only Speech Phones - $GOP_{Speech}$

Alternatively, the utterance level  $GOP$  score can be calculated using only speech phones. This excludes all non-speech phones - those forming part of silence or noise. It is argued that models for such non-speech sounds are often poorly focussed, which could lead to a severe mismatch between force aligned and free phone loop recognised phones and result in relatively high  $GOP$  scores. This could adversely affect an utterance level score aimed at rating pronunciation. This  $GOP$  score variant will be referred to as  $GOP_{Speech}$ .

### 5.1.3 Only Phones in the Context of Speech Phones - $GOP_{Context}$

Utterance level  $GOP$  scores can be refined further to  $GOP_{Context}$  by excluding non-speech phones as well as speech-phones in either the left or right context of non-speech phones. Because models for non-speech sounds are often poorly focussed, the resulting alignments of such phones are often inaccurate. These alignment errors also affect the phones in the left and right context of the non-speech sounds. The concern is that these alignment issues could adversely affect the  $GOP$  scores of phones bordering on non-speech sounds, due to the resulting mismatch between force-aligned and free phone loop phones.

### 5.1.4 Word Level Normalisation - $GOP_{WordLvl}$

$GOP_{WordLvl}$  refers to utterance level  $GOP$  scores that are time normalised on the word level rather than the phone level. A similar score was investigated in [19]. Each phone's  $GOP$  score is weighted by the phone's duration, summed over all phones in a word, and normalised by the word's duration. This allows the  $GOP$  scores of longer phones to have a greater effect on the word score. In this case the  $GOP$  score for a word consisting of  $M$  phones is defined as:

$$GOP(word_j) = \frac{\sum_{i=1}^M GOP(q_i) \cdot d(q_i)}{\sum_{i=1}^M d(q_i)}$$

where  $d(q_i)$  denotes the durations of phone  $q_i$ . The utterance level  $GOP$  is then calculated by summing the  $GOP$  scores of all the words in the utterance, and normalising by the number of words,  $W$ :

$$GOP_{WordLvl} = \frac{\sum_{j=1}^W GOP(word_j)}{W}$$

## 5.2 Results

To evaluate the potential of the  $GOP$  score variations presented above to predict human assessments of oral proficiency, their correlations with human ratings are calculated. We are most interested in the correlation with human assessments of pronunciation, as the  $GOP$  score is an acoustic (rather than a temporal) measure, intended specifically to determine pronunciation quality [6].

Based on the  $GOP$  algorithm, we expect high  $GOP$  scores for poorly pronounced phones, and low  $GOP$  scores for well pronounced phones. In the best case, when the free phone loop recognition results in exactly the same segmentation and phone sequence as the force aligned recognition, the  $GOP$  score will be 0. Given the definition of the human rating scales used in this study, where lower ratings indicate higher proficiency, we expect  $GOP$  to have a positive correlation with human proficiency ratings.

For each variant of the *GOP* algorithm, scores are determined using each of the four forced alignment recognition strategies described in Section 4.2. Two of these recognition strategies, the use of a finite state grammar (FSG) and the use of a unigram language model, can be implemented automatically, and are used for the reading task and repeating task respectively. We are most interested in the correlations resulting from these automatic recognition strategies. Correlations based on the use of an oracle FSG give an indication of the effects better recognition accuracy would have on the score’s performance. The correlations calculated using oracle alignment can be seen as a best-case-scenario.

### 5.2.1 $GOP_{All}$

Table 5.1 shows the correlations of  $GOP_{All}$  scores with human ratings. It shows that, for our data,  $GOP_{All}$  is very poorly correlated with human ratings for *Pronunciation*, contrary to expectation.

The correlation of  $GOP_{All}$  scores with ratings in the three reading task scales, *Hesitation*, *Pronunciation* and *Intonation*, are negligible, regardless of the recognition strategy used. For the two repeating task scales, *Success* and *Accuracy*, there is some correlation. For these scales, the scores calculated using a unigram language model resulted in a higher correlation than when an oracle FSG grammar was used.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	0.05	0.02	0.07		
Unigram				0.39	0.35
Oracle FSG	0.06	-0.04	0.05	0.28	0.30
Oracle	0.04	0.03	-0.04		

**Table 5.1:** Correlation of  $GOP_{All}$  scores with human ratings for different rating scales and recognition strategies.

### 5.2.2 $GOP_{Speech}$

The correlations of  $GOP_{Speech}$  scores with human ratings are shown in Table 5.2. It can be seen that using only speech phones to calculate the utterance level *GOP* did not result in any clear improvement on the performance of  $GOP_{All}$  with regard to the reading task. The correlations with both repeating task scales are slightly higher for both of the recognition strategies applied to the repeating task than they were for  $GOP_{All}$ . This seems to indicate that where correlation with human ratings exists, this correlation can be improved by using only speech phones to calculate the utterance level *GOP* score.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	0.06	0.00	0.07		
Unigram				0.42	0.39
Oracle FSG	0.08	-0.04	0.06	0.31	0.33
Oracle	0.11	0.10	0.05		

**Table 5.2:** Correlation of  $GOP_{Speech}$  scores with human ratings for different rating scales and recognition strategies.

### 5.2.3 $GOP_{Context}$

Table 5.3 shows the correlation of  $GOP_{Context}$  scores with human ratings. Correlations with the three reading task scales are still negligible, except where the  $GOP$  scores are calculated using oracle recognition. This correlation suggests that  $GOP$  scores are somewhat correlated with human ratings for the reading task, but not sufficiently to be used effectively for our data under real world circumstances. It also suggests that improved accuracy of the automatic recogniser might lead to higher correlation between the  $GOP_{Context}$  scores and human ratings.

With regard to the two repeating task scales, the correlations of  $GOP_{Context}$  with these human ratings are consistently higher than those of  $GOP_{All}$  and  $GOP_{Speech}$ . These are the highest correlations with human ratings achieved by any of the  $GOP$  variants investigated here.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	0.08	-0.02	0.08		
Unigram				0.45	0.41
Oracle FSG	0.10	-0.04	0.06	0.33	0.33
Oracle	0.22	0.23	0.16		

**Table 5.3:** Correlation of  $GOP_{Context}$  scores with human ratings for different rating scales and recognition strategies.

### 5.2.4 $GOP_{WordLvl}$

The correlations of human ratings with  $GOP_{WordLvl}$  scores are shown in Table 5.4. The negative correlations with human ratings for the reading task show that normalising  $GOP$  scores on the word level does not improve the ability of the  $GOP$  algorithm to predict human assessments. Given the definition of our rating scales and the  $GOP_{WordLvl}$  score, one would expect positive correlations. The correlations with human ratings for the repeating task are lower than those of the other  $GOP$  variants investigated in this study.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	-0.07	-0.14	-0.08		
Unigram				0.31	0.25
Oracle FSG	-0.07	-0.19	-0.13	0.25	0.27
Oracle	-0.10	-0.19	-0.13		

**Table 5.4:** *Correlation of  $GOP_{WordLvl}$  scores with human ratings for different rating scales and recognition strategies.*

### 5.3 Limiting Phone Scores and Phone Durations

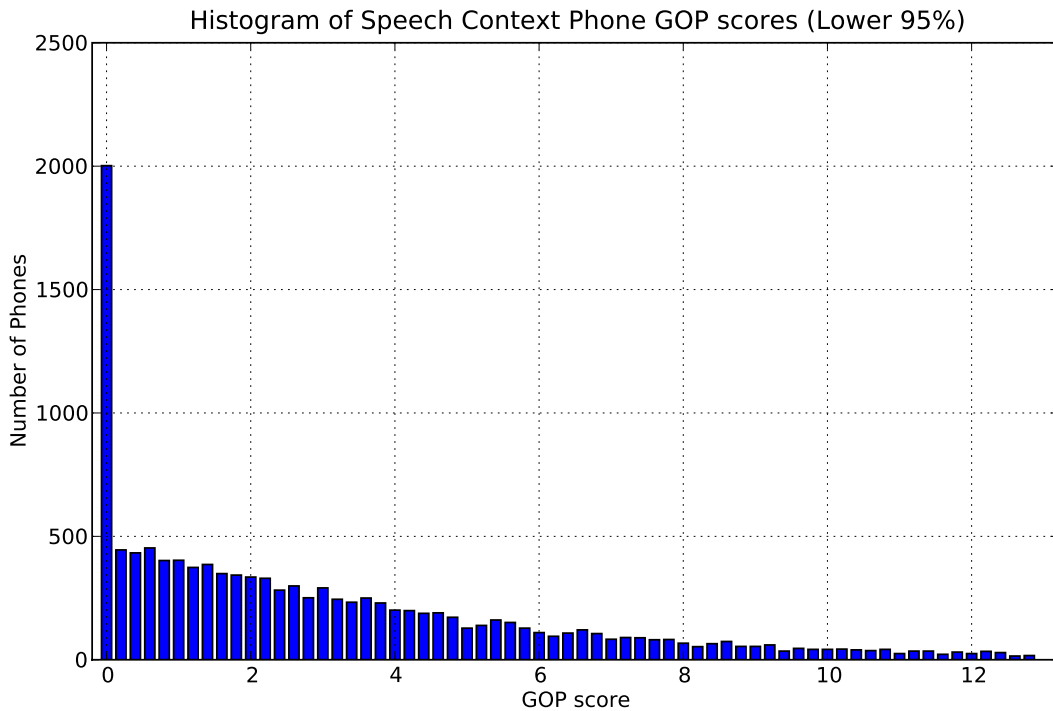
The correlations between  $GOP$  scores and human ratings of pronunciation obtained in this chapter are lower than those found by other authors using similar reading tasks [6; 7; 9]. Although the differences in rating scales and test dataset sizes between research groups make direct comparison of correlation values difficult, the extent of the discrepancy is surprising. To investigate the factors affecting the performance of  $GOP$  scores in more detail, we have examined the distribution of phone durations and the associated phone level  $GOP$  scores. The aim was to investigate whether poor results obtained for either very short or for very long phones were affecting the overall success.

Of the  $GOP$  variations calculated,  $GOP_{Context}$  has shown the most promise, and based on the work by Witt & Young [6] and the fact that the  $GOP$  score is an acoustic measure, higher correlations are most expected with *Pronunciation* ratings. Hence we focus on the correlation between  $GOP_{Context}$  and *Pronunciation* ratings.

Figure 5.2 shows the distribution of phone level  $GOP$  scores assigned to the 12616 phones in speech context present in the FSG recognition of the reading task. When free phone loop recognition and forced alignment recognition recognise the same phone over the same frames, the resulting  $GOP$  score for that phone is 0. This is the cause of the high occurrence of the score 0.

The histogram in Figure 5.2 has been truncated, showing only the lower 95% of assigned  $GOP$  scores. Of the assigned phone level  $GOP$  scores, 90% are below 9.32. However, some phones were assigned scores as high as 63.10. Such high scores are most likely the result of severe alignment errors, and can have a significant effect on the  $GOP$  score of the utterance. To investigate the effect of these very high phone level  $GOP$  scores, we considered three separate sets of utterance level  $GOP$  scores:

- For the first set, utterance level scores are calculated using all available speech context phones, irrespective of the  $GOP$  scores assigned to them. This is the  $GOP_{Context}$  score as determined before.



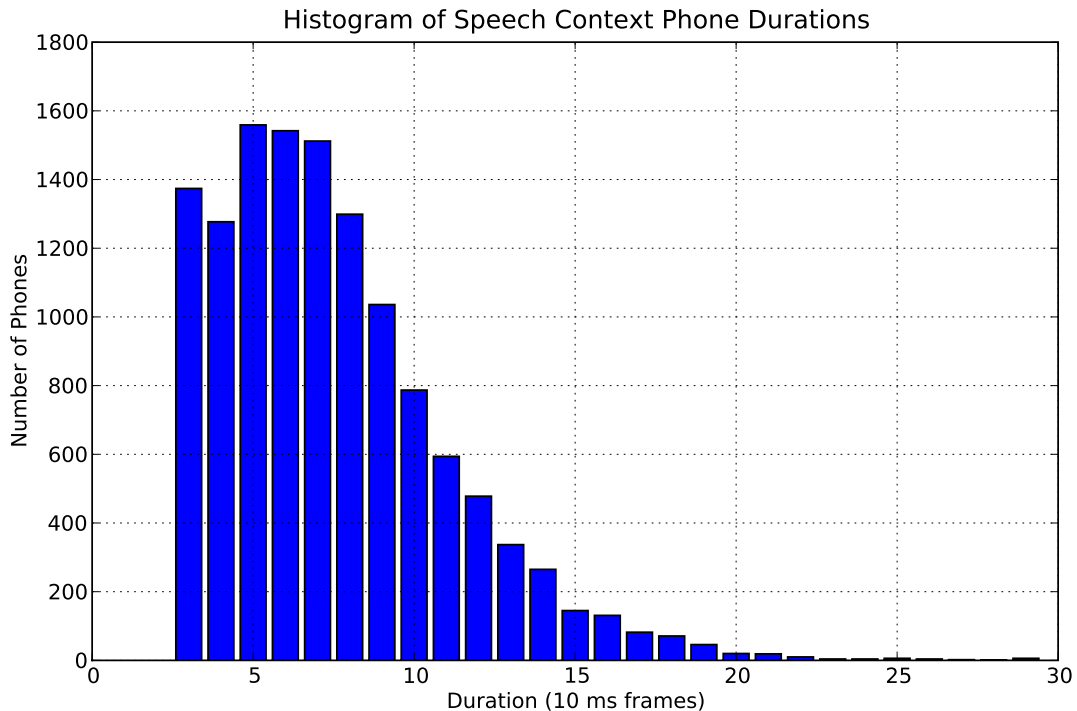
**Figure 5.2:** *Distribution of phone level GOP scores assigned to phones in speech context for the reading task.*

- For the second set, utterance level scores are calculated using a maximum *GOP* score of 13. This allows 95% of *GOP* scores to be used as is, while the other 5%, with scores above 13, are replaced by the maximum score of 13. This strategy aims to mitigate the effect of phones with very high scores.
- For the third set, utterance level scores are calculated using a maximum *GOP* score of 9.32. This allows 90% of *GOP* scores to be used as is, while the other 10%, with scores above 9.32, are replaced by the maximum score of 9.32. This approach also seeks to mitigate the effect of phones with very high scores.

Figure 5.3 shows the distribution of phone durations, in frames, for the 12616 speech context phones recorded in the reading task. Phones must be at least 3 frames long, since 3-state HMM acoustic models are used.

It was suspected that longer or shorter phones may be more or less suitable for calculating an utterance level *GOP* score. One could investigate this by calculating separate sets of utterance level *GOP* scores based on each occurring phone duration, and correlating these scores with the associated human ratings. However, our dataset is too small to allow analysis in such level of detail. Especially for longer durations, there are too few phones to calculate reliable *GOP* scores.

Instead, we calculate sets of utterance level *GOP* scores by imposing an upper duration



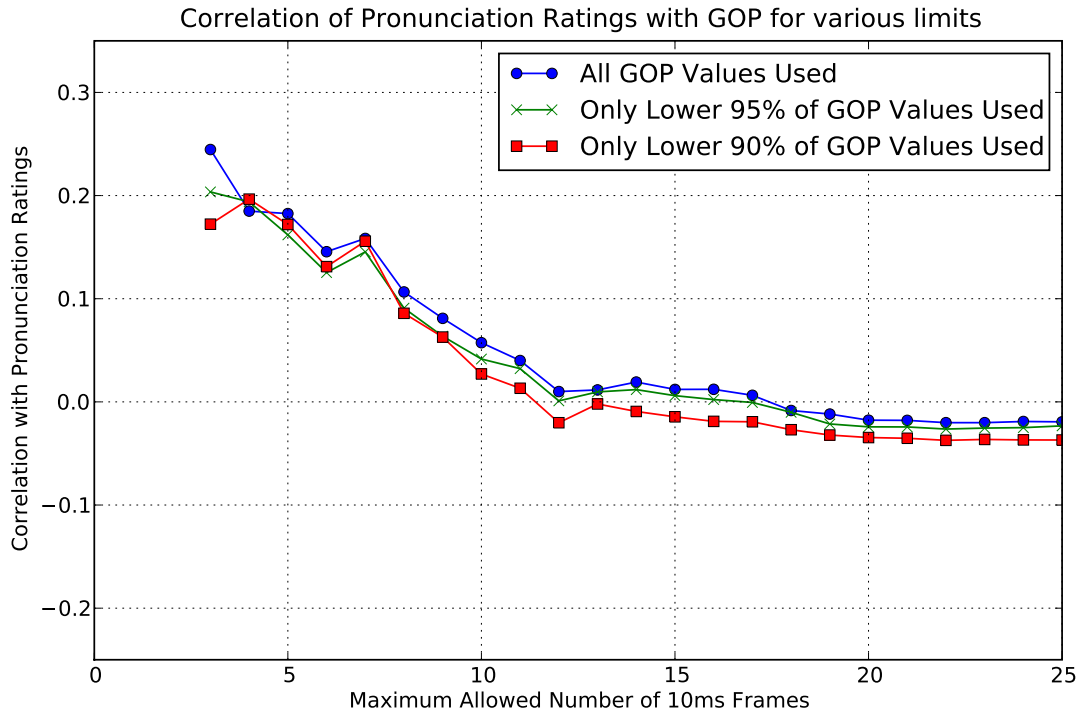
**Figure 5.3:** *Distribution of durations of phones in speech context for the reading task.*

limit on each set. Thus, for the set corresponding to a maximum duration of 3 frames, only the phones consisting of 3 frames are used to calculate utterance level scores. For the set corresponding to a maximum duration of 4 frames, phones consisting of 3 frames as well as phones consisting of 4 frames are used. As the upper duration limit is increased, more phones are included in the utterance level calculations. For the set corresponding to 25 frames, all phones with a duration up to and including 25 frames are used, constituting 99.8% of the total available phones in the reading task.

For each of the three *GOP* score limits described above ( $\infty$ , 13 and 9.32), 23 sets of utterance level scores are calculated, one for each of the upper duration limits from 3 frames to 25 frames. The correlations of these sets with the human ratings for *Pronunciation* are shown in Figure 5.4. Each line corresponds to one of the *GOP* score limits and shows the trend of correlation values for that score limit as the duration limit is increased.

From Figure 5.4 it is apparent that excluding very high *GOP* scores by limiting the maximum allowed phone score makes no great difference to the correlation with human ratings. The figure also shows that the correlation with human pronunciation ratings is the highest when only the shortest available phones, those consisting of 3 frames, are used for calculating utterance level scores. As the maximum allowed phone duration increases, the correlation with human ratings decreases. Although the correlations are not very high, even in the best case (0.24), this trend is investigated further in the following.

In order to assess the robustness of the trend, the dataset was split into two subsets ( $A$



**Figure 5.4:** *Correlations of Pronunciation ratings with  $GOP_{Context}$  scores against allowed maximum phone duration.*

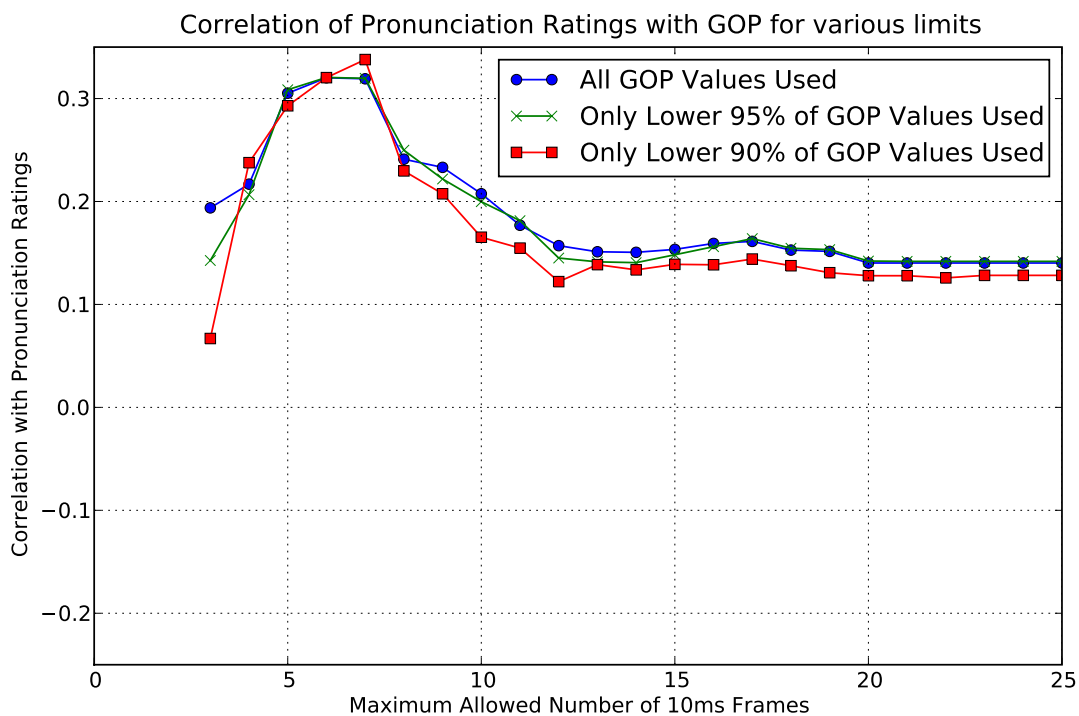
and *B*). Each subset contained 45 students, chosen in such a way that the two sets had approximately the same average *Pronunciation* rating. There was no overlap of students between the two subsets.

The correlations with *Pronunciation* ratings were then calculated for each of the 23 duration limits and each of the 3 *GOP* score limits for both subset *A* and subset *B*. The results are shown in Figures 5.5 and 5.6 respectively.

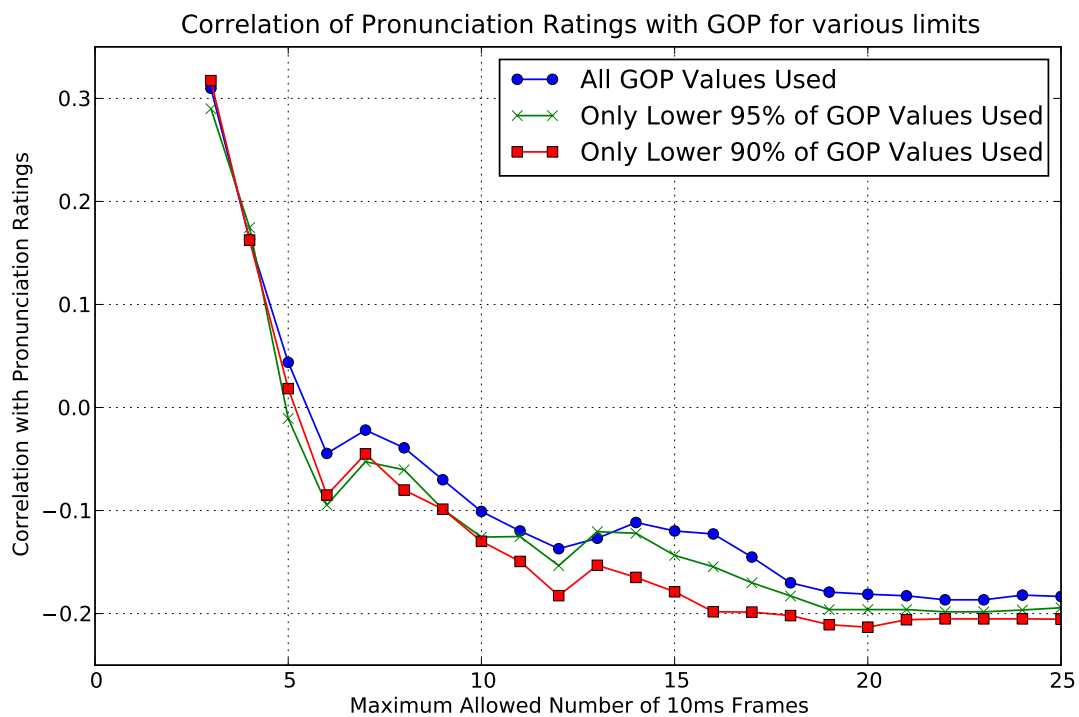
Comparing Figure 5.4 with Figures 5.5 and 5.6, our finding that limiting the maximum *GOP* score makes little difference to the correlation with pronunciation ratings is confirmed.

The full set and the two subsets respond in different ways to the increase of allowed phone duration. We believe this is due to the relatively small size of the sets. In all three cases, however, the *GOP* scores of shorter phones seem to be better correlated with *Pronunciation* ratings than the *GOP* scores of longer phones. It is possible that an automatic rating system could benefit by taking this effect into consideration when calculating utterance level scores, for example by assigning different weights to the *GOP* scores of phones based on their durations. However, developing such a system of weights would require a data corpus large enough to contain representative samples of each relevant phone duration. The corpus currently available for this research does not allow for such an investigation.





**Figure 5.5:** Subset A: *Correlations of Pronunciation ratings with  $GOP_{Context}$  scores against allowed maximum phone duration.*



**Figure 5.6:** Subset B: *Correlations of Pronunciation ratings with  $GOP_{Context}$  scores against allowed maximum phone duration.*

## 5.4 Summary and Conclusions

We determined the correlations between four variations of the posterior log-likelihood score *Goodness of Pronunciation* and associated human ratings for a set of test utterances. The highest correlations were found using  $GOP_{Context}$ , a *GOP* score that excludes any non-speech phones and phones adjacent to non-speech phones when calculating the utterance level score.

In general, the correlation values for the reading task scales were negligible. There are indications that better recognition accuracy may lead to better results. The fact that most students were given very high ratings for the reading task and the associated small variation in ratings may also be a contributing factor. This corresponds to the findings of Zechner et al., who has also concluded that testing students concentrated in the high-proficiency end of the rating scale leads to lower correlations with machine scores [15].

The correlations between the *GOP* scores and ratings for the repeating task scales ranged from 0.25 to 0.45. On their own these *GOP* scores may not predict human proficiency ratings with sufficient accuracy. However, they may prove useful when combined with other machine scores. Combinations of machine scores are investigated in Chapter 8.

The effect of limiting the maximum allowed phone level *GOP* score value was investigated for the reading task. No great difference in correlation with *Pronunciation* ratings was found. The effect of limiting the maximum allowed phone duration was also investigated. The exclusion of longer phones from the utterance level *GOP* scores lead to better correlations with human ratings for *Pronunciation*. Further research into duration-specific *GOP* scores using more extensive datasets could prove useful in finding ways of improving the correlation between posterior log-likelihood scores and human proficiency ratings.

# Chapter 6

## Scores Based On Segmentation

This chapter presents algorithms which calculate machine scores based on the segmentation of utterances. These scores focus on the *temporal* features of speech, rather than on its acoustic characteristics.

During the recognition process, the Viterbi algorithm aligns phones with audio segments, estimating where one phone ends and another begins. The scoring algorithms in this chapter rely on this phone level alignment to determine scores for the utterances to be evaluated.

We will describe four different segmentation based scoring algorithms, and evaluate their potential for predicting human assessments of oral proficiency by calculating the correlations of the machine scores with the human rating scales applied to our data.

### 6.1 Score Definitions

After processing an utterance, the recognition output of the HTK *HVite* function contains a list of recognised phones and their start and end times, describing the duration of individual phones. These phones can be classified as *speech phones*, those forming part of words, and *non-speech phones*, those forming part of silence or noise.

#### 6.1.1 *Rate of Speech*

The *Rate of Speech* (*ROS*) of an utterance is defined in [8] as the number of speech phones per second, calculated using the number of speech phones in the utterance  $M_{Speech}$ , and the total duration of the utterance  $T_{Total}$ , in seconds:

$$ROS = \frac{M_{Speech}}{T_{Total}}$$

Any silences leading or trailing the utterance are ignored when determining the total duration.

### 6.1.2 *Articulation Rate*

*Articulation Rate* (*ART*) is similar to *ROS*, but does not take the duration of silence and noise in the utterance into account [13]. It is calculated using the total duration of speech phones in the utterance,  $T_{Speech}$ , rather than the total duration:

$$ART = \frac{M_{Speech}}{T_{Speech}}$$

### 6.1.3 *Phonation/Time Ratio*

The *Phonation/Time Ratio* (*PTR*) is the fraction of the utterance duration that consists of speech phones [12]. It is defined as:

$$PTR = \frac{T_{Speech}}{T_{Total}}$$

where  $T_{Speech}$  is the duration of all speech phones in the utterance and  $T_{Total}$  is the total duration of the utterance, ignoring leading or trailing silences.

### 6.1.4 *Segment Duration Score*

The *Segment Duration Score* (*SDS*) compares the duration of each phone in an utterance with the expected duration of that phone based on training data. It is based on the argument that the training data reflects the pronunciation expected from proficient speakers [7].

To allow for variations in speech rate between speakers, the duration of each phone is normalised by multiplication with the utterance *ROS*. This is done for the utterances to be evaluated, as well as the training data. We define this normalised duration of a phone  $q_i$  as  $f(q_i)$ , where  $d_i$  is the duration of  $q_i$ :

$$f(q_i) = d_i \cdot ROS$$

The phone level *SDS* is defined as the probability of the normalised duration of the phone, given the type of phone:

$$SDS(q_i) = C_{q_i} \cdot p(f(q_i)|q_i)$$

The probability  $p(f(q_i)|q_i)$  is based on a discrete distribution of normalised durations for the given phone, determined from the training data. The scaling factor  $C_{q_i}$  is associated with the probability distribution for the given phone, and is defined so that a phone duration corresponding to the peak of the probability distribution results in a phone level *SDS* score of 1.

Using the probability  $p(f(q_i)|q_i)$  as a score without scaling would result in uneven scoring between different phones. Lower scores would be assigned to phones with broader probability distributions, which do not have well-defined peaks, even when pronounced perfectly, i.e. with a normalised duration matching that of the distribution peak. By associating a scaling

factor with each probability distribution, we can scale scores in such a way that all perfectly pronounced phones are assigned the same maximum score of 1, irrespective of the shape of their duration distributions.

Finally, the utterance level *SDS* is defined as the average phone level *SDS* for the given utterance. For reasons discussed below, only 34 of the 52 monophones used by the recogniser can be assigned *SDS* scores. The utterance *SDS* is based on all instances of these 34 phones in the utterance.

It could be argued that speech phones in the left or right context of non-speech phones should be excluded from the calculation of utterance level *SDS* scores, as these phones are often poorly aligned. However, preliminary investigation showed this to result in consistently lower correlations between the *SDS* scores and human assessments for our data, and it is therefore not considered further in this study.

## Distributions

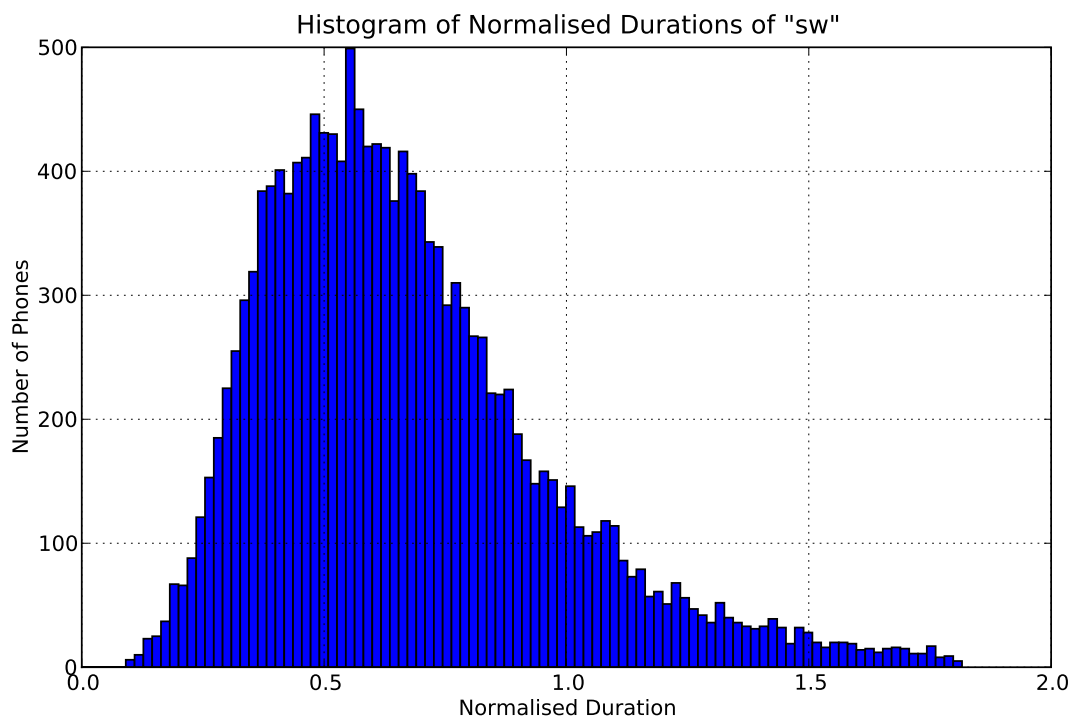
To calculate the probability  $p(f(q_i)|q_i)$ , a discrete probability distribution of normalised durations is required for each monophone. These probability distributions are based on training data.

The recogniser is capable of recognising 52 different monophones. Not all of these phones are sufficiently represented in the training data to allow the calculation of an accurate probability distribution. We therefore base the *SDS* on the 34 monophones that occur most frequently in the training data. These 34 monophones make up approximately 95% of the phones in the training data.

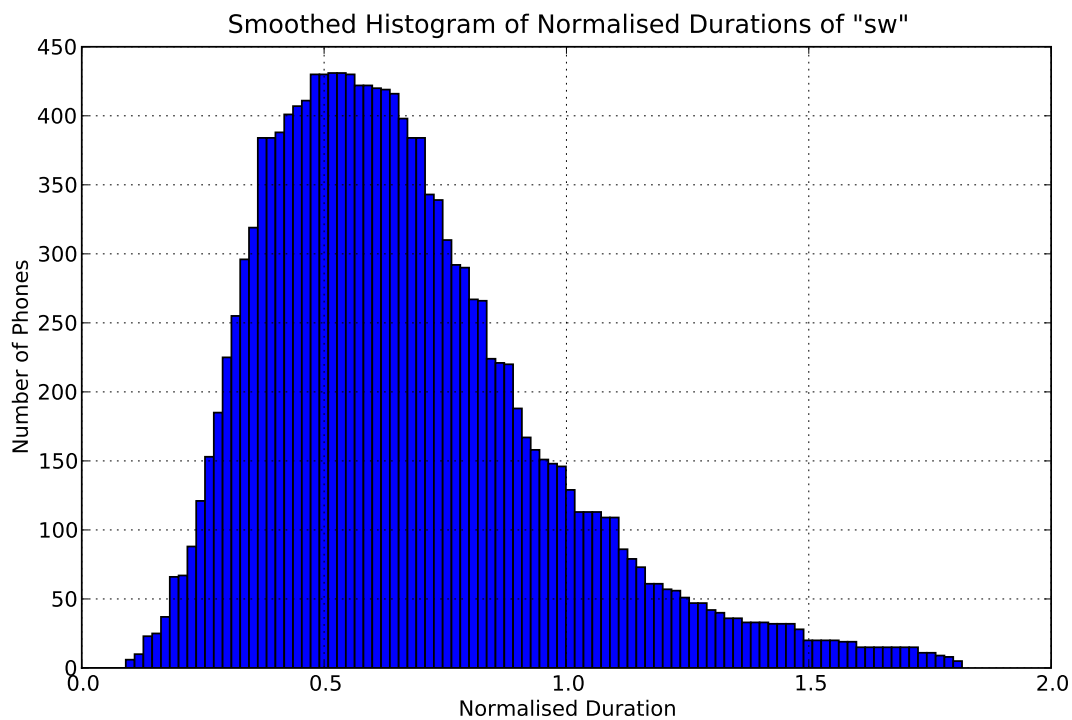
For each of the 34 monophones a histogram of normalised durations was determined, using 100 bins of equal width. To eliminate outliers, only the lower 99% of normalised durations were used for each histogram. The number of available normalised durations per monophone ranged between 17202 and 1602. Each histogram was smoothed using a median filter with a window size of 5. Finally, each monophone’s histogram of normalised durations was scaled so the bin heights sum to 1. This was then used as the discrete probability distribution of normalised durations for the given monophone. The distributions used for the 34 monophones are shown in Appendix C.

Using the monophone “*sw*” as an example, Figures 6.1, 6.2 and 6.3 show the corresponding histogram, smoothed histogram and probability distribution respectively. The histogram is based on 15596 normalised durations taken from the training data. The scaling factor associated with “*sw*”,  $C_{sw}$ , is 35.587, calculated using the peak value of the probability distribution:

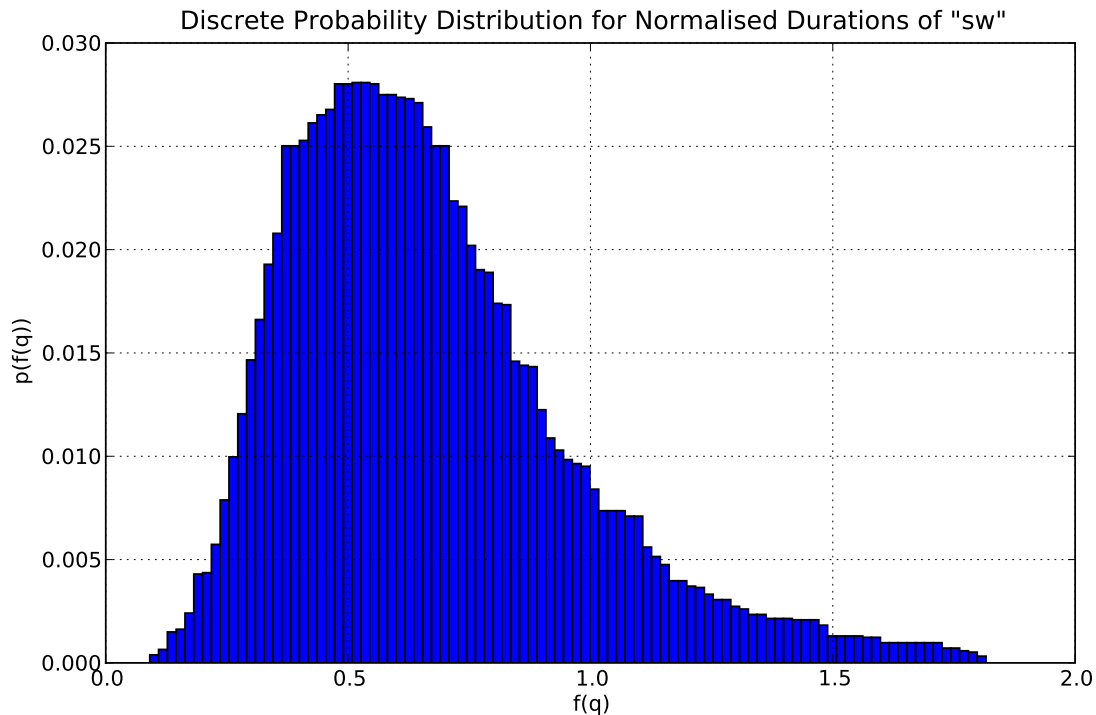
$$C_{sw} = \frac{1}{Peak_{sw}} = \frac{1}{0.0281} = 35.587$$



**Figure 6.1:** *Histogram of normalised durations of the phone “sw” based on training data.*



**Figure 6.2:** *Histogram of normalised durations of the phone “sw” after smoothing with median filter.*



**Figure 6.3:** *Discrete probability distribution of normalised duration of the phone “sw” based on training data.*

## 6.2 Results

We evaluate the potential of the four segmentation based scores introduced in this chapter to predict oral proficiency assessments by calculating their correlations with human ratings. When comparing the performance of the different scores with each other, we pay special attention to the two automatically realisable recognition strategies, the use of a finite state grammar (FSG) (for the reading task) and the use of a unigram language model (for the repeating task). The two other recognition strategies, the use of an oracle FSG and oracle recognition, require human transcriptions of the utterances to be rated. The recognition strategies are discussed in detail in Section 4.2.

All four of the scoring algorithms are expected to result in high scores for high proficiency and low scores for low proficiency. Given the definitions of our human rating scales, which assign lower values to higher proficiency, we expect negative correlations between the segmentation scores and human ratings.

### 6.2.1 *Rate of Speech*

The *Rate of Speech* scores are relatively well correlated with all human rating scales, as shown in Table 6.1.

Based on the two automatic recognition strategies, *ROS* has the highest correlations of all segmentation based scores with the human ratings for *Intonation*, *Success* and *Accuracy*.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	Succ.	Acc.
FSG	-0.54	-0.48	<b>-0.49</b>		
Unigram				<b>-0.67</b>	<b>-0.65</b>
Oracle FSG	-0.50	-0.50	-0.50	-0.65	-0.62
Oracle	-0.94	-0.77	-0.95		

**Table 6.1:** *Correlation of Rate of Speech scores with human ratings for different rating scales and recognition strategies.*

### 6.2.2 *Articulation Rate*

Table 6.2 shows the correlations of the *Articulation Rate* scores with the various human rating scales. The performance of *ART* is similar to that of *ROS*, with a slight increase in correlation with the ratings for *Pronunciation*. The correlation with *Pronunciation* ratings is the highest correlation with these ratings of all segmentation scores based on automatic recognition.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	-0.41	<b>-0.50</b>	-0.46		
Unigram				-0.60	-0.58
Oracle FSG	-0.40	-0.52	-0.45	-0.30	-0.32
Oracle	-0.87	-0.86	-0.88		

**Table 6.2:** *Correlation of Articulation Rate scores with human ratings for different rating scales and recognition strategies.*

### 6.2.3 *Phonation/Time Ratio*

The correlations of *Phonation/Time Ratio* scores with the various human rating scales are shown in Table 6.3. *PTR* is best correlated with the human ratings for *Hesitation*, and presents the highest correlation with this scale of all the segmentation scores, based on an automatic recognition strategy. The correlations between *PTR* scores and the other human rating scales are low compared to those of *ROS* and *ART*.



	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	<b>-0.64</b>	-0.18	-0.39		
Unigram				-0.45	-0.44
Oracle FSG	-0.50	-0.17	-0.34	-0.70	-0.63
Oracle	-0.31	0.09	-0.33		

**Table 6.3:** *Correlation of Phonation/Time Ratio scores with human ratings for different rating scales and recognition strategies.*

### 6.2.4 Segment Duration Score

Table 6.4 shows the correlations of the *Segment Duration Score* scores with the different human rating scales. The correlations with the three reading task scales are negligible, except when using oracle recognition. The correlations with the two repeating task scales are comparable with those of *ART*, but lower than those of *ROS*.

	<i>Hesit.</i>	<i>Pronun.</i>	<i>Inton.</i>	<i>Succ.</i>	<i>Acc.</i>
FSG	0.15	-0.18	0.00		
Unigram				-0.61	-0.56
Oracle FSG	0.11	-0.17	0.06	-0.48	-0.47
Oracle	-0.46	-0.54	-0.43		

**Table 6.4:** *Correlation of Segment Duration Score scores with human ratings for different rating scales and recognition strategies.*

## 6.3 Summary and Conclusions

In general, the segmentation scores discussed in this chapter are better correlated with human ratings than the likelihood scores of Chapter 5.

*Rate of Speech* scores are arguably the most simple to calculate, and correlate well with all human rating scales. *Articulation Rate*, which is closely related to *ROS* did not perform better, except for a slight improvement in correlation with *Pronunciation* ratings. The *Phonation/Time Ratio* scores show promise as a predictor of human ratings for *Hesitation* in particular. The *Segment Duration Score* scores have no usable correlation with human ratings for the reading task, but are relatively well correlated with ratings for the repeating task.

The effects of combining these and other scores to predict human proficiency ratings more accurately are investigated in Chapter 8.

# Chapter 7

## Scores Based On Repeat Accuracy

This chapter presents algorithms aimed at automatically assessing the accuracy of student responses. We will describe three accuracy scoring algorithms and present their correlations with the human rating scales for the repeating task, namely *Success* and *Accuracy*.

Since it is assumed that the reading task prompts were read without error, we apply the accuracy scoring algorithms only to the repeating task.

### 7.1 Score Definitions

For each prompt in the repeating task there is a correct response, or desired utterance. All of the accuracy algorithms presented here are based on a comparison of the recogniser output with an orthographic reference transcription of the desired utterance.

#### 7.1.1 *HResults Accuracy*

This score is calculated using the HTK tool *HResults*, which uses a dynamic programming-based string alignment procedure to align the recogniser output with the reference transcription [17]. It counts the number of correctly aligned words ( $H$ ), the number of insertions ( $I$ ), the number of deletions ( $D$ ), and the number of words in the reference transcription ( $W$ ). The *HResults Accuracy* ( $Acc_{HResults}$ ) is then calculated as:

$$Acc_{HResults} = \frac{H - I}{W} \times 100\%$$

Note that this score is penalised by insertions. When the number of insertions exceeds the number of correctly recognised words, the score is negative.

#### 7.1.2 *HResults Correct*

The *HResults Correct* ( $Cor_{HResults}$ ) score indicates the percentage of reference transcription words present in the recogniser output [17]. In contrast to  $Acc_{HResults}$ , this score does not take insertions into account. It is calculated by the HTK tool *HResults*, as described above, and is defined as:

$$Cor_{HResults} = \frac{H}{W} \times 100\%$$

### 7.1.3 *Weighted Correct*

When calculating  $Cor_{HResults}$ , all words in the reference transcription are regarded as equally important. However, it is plausible that when human raters are assigning values to *Accuracy* or *Success*, they may penalise speakers less for missing certain unimportant words than for missing words that are more central to the semantic meaning of the target utterance.

To investigate this, we assign a rank to each word in the reference transcription as a measure of that word’s semantic importance. In some cases, adjacent words are grouped together to form a phrase, which is then assigned a single rank. Each rank is associated with a *weight*, which represents the number of marks that will be awarded if the corresponding word or phrase occurs in the recogniser output.

The *Weighted Correct* ( $Cor_{Weighted}$ ) score is defined as the percentage of marks that were awarded:

$$Cor_{Weighted} = \frac{\sum_{i=1}^H w_i}{\sum_{j=1}^W w_j} \times 100\%$$

where  $H$  is the number of correct words or phrases and  $W$  is the total number of ranked words or phrases in the reference transcription.  $w_i$  is the weight associated with the  $i^{th}$  correct word or phrase, and  $w_j$  is the weight associated with the  $j^{th}$  word or phrase in the reference transcription.

The eight prompts for the repeating task were analysed<sup>1</sup> and a rank from 1 to 6 was assigned to each word or semantic group of words. Ranks were assigned by identifying the *head* of the sentence and elements that modify the head, as defined in [20].

A rank of 1 was assigned to words or phrases with the least semantic importance, and a rank of 6 to those with the most semantic importance. The prompts together with their word and phrase ranks are shown in Appendix B.

## Weights

The weights associated with the different ranks can be adjusted in an effort to approximate the relative importance human raters would attach to each rank. The more accurate the approximation, the stronger the correlation between the  $Cor_{Weighted}$  scores and human ratings should be.

We investigate four sets of weights based on four different mathematical relationships between the ranks. Where  $w_r$  is the weight associated with rank  $r$ , the four sets of weights

---

<sup>1</sup>Personal communication with Prof. C. van der Walt, Department of Curriculum Studies, Faculty of Education, Stellenbosch University, who designed the prompts for the automated test.

are defined as:

Equal:	$w_r = 1$
Linear:	$w_r = r$
Quadratic:	$w_r = r^2$
Logarithmic:	$w_r = \log(r) + 1$

Ranks are numbered 1 to 6. A constant of 1 is added to the logarithmic weights to avoid a weight of 0 for the lowest rank. Table 7.1 shows the weight sets used, calculated based on the above equations.

Weight Set	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6
<i>Equal</i>	1	1	1	1	1	1
<i>Linear</i>	1	2	3	4	5	6
<i>Quadratic</i>	1	4	9	16	25	36
<i>Logarithmic</i>	1.00	1.69	2.10	2.39	2.61	2.79

**Table 7.1:** Four different weight sets associated with the Weighted Correct ranks.

## 7.2 Results

To evaluate the correspondence between the three accuracy based scoring algorithms and the human ratings for the repeating task, we calculate the correlations between the various scores and the human ratings.

Results for two recognition strategies are presented. While recognition using a unigram language model can be automated, the use of an oracle FSG grammar requires human transcriptions of the utterance to be recognised. The results for an oracle FSG grammar are included here to indicate what the effect of better recognition accuracy would be on the performance of the machine scores. The recognition strategies are discussed in detail in Section 4.2.

The three scoring algorithms are all expected to result in high scores for high proficiency and low scores for low proficiency. Given the definition of the human rating scales, which assign lower values to higher proficiency, we expect negative correlations between the accuracy scores and human ratings.

### 7.2.1 *HResults Accuracy*

Table 7.2 shows the correlation of the *HResults Accuracy* scores with human ratings for the repeating task. When based on recognition performed using a unigram language model,

the correlations are comparable with those obtained for *Rate of Speech* (Table 6.1). The results based on recognition using an oracle FSG grammar show that better recognition may improve the correlation values.

	<i>Success</i>	<i>Accuracy</i>
Unigram	-0.61	-0.63
Oracle FSG	-0.81	-0.77

**Table 7.2:** *Correlation of HResults Accuracy scores with human ratings for different rating scales and recognition strategies.*

### 7.2.2 *HResults Correct*

The *HResults Correct* scores have high correlations with human ratings, as shown in Table 7.3. These are the highest correlations between a machine score and human ratings for *Success* and *Accuracy* found in this study. The results based on recognition using an oracle FSG grammar show that higher recognition accuracy may lead to even better correlations between the machine scores and the human ratings.

	<i>Success</i>	<i>Accuracy</i>
Unigram	-0.76	-0.85
Oracle FSG	-0.87	-0.90

**Table 7.3:** *Correlation of HResults Correct scores with human ratings for different rating scales and recognition strategies.*

### 7.2.3 *Weighted Correct*

Table 7.4 shows the correlations between human ratings and *Weighted Correct* scores based on four different weight sets described in Section 7.1.3. While better than those obtained for *HResults Accuracy*, the correlations are lower than those found for *HResults Correct*, which regards all words as equally important.

The weight sets with equal weights and logarithmically related weights resulted in the highest correlations. When using equal weights, the *Weighted Correct* score is similar to the *HResults Correct* score except for the grouping of some words into phrases.

As with *HResults Accuracy* and *HResults Correct*, the correlations are higher when based on recognition using an oracle FSG grammar, indicating that better recognition may improve results.

Weight Set	<i>Success</i>		<i>Accuracy</i>	
	Unigram	Oracle FSG	Unigram	Oracle FSG
<i>Equal</i>	<b>-0.71</b>	-0.86	<b>-0.79</b>	-0.90
<i>Linear</i>	-0.62	-0.80	-0.73	-0.84
<i>Quadratic</i>	-0.47	-0.68	-0.59	-0.71
<i>Logarithmic</i>	<b>-0.70</b>	-0.85	<b>-0.79</b>	-0.89

**Table 7.4:** *Correlation of Weighted Correct scores with human ratings for different rating scales, recognition strategies and weight sets.*

### Random Weights

In an attempt to find an approximately optimal weight set, correlations were calculated for 4000 weight sets made up of randomly-drawn values. These were uniformly distributed integers between 0 and 100. Groups of six random values were generated and sorted numerically to form weight sets.

Based on these 4000 random weight sets, correlations with human ratings for *Success* ranged between 0.29 and 0.73, and correlations with the *Accuracy* scale ranged between 0.40 and 0.81. Hence no better alternative to the straightforward application of *HResults Correct* was found.

## 7.3 Summary and Conclusions

In comparison with the posterior log-likelihood scores of Chapter 5 and the segmentation based scores of Chapter 6, each of the three scores presented in this chapter correlates well with human ratings.

The fact that *HResults Correct* performed better than *HResults Accuracy* leads us to suspect that human raters do not penalise students for word insertions in the way *HResults Accuracy* does. Instead, they appear to focus primarily on the number of correct words.

Even the highest correlation achieved by the *Weighted Correct* score, that of 0.81 with a random weight set, is not as high as that of *HResults Correct*, 0.85. Of the mathematically calculated weight sets, those with equal weights and logarithmically related weights resulted in the highest correlations. This shows that the best results are achieved when *Weighted Correct* is most similar to *HResults Correct*, and leads us to believe that manually ranking words and phrases in a target utterance according to their semantic importance is not a promising method of improving automatic accuracy scoring.

# Chapter 8

## Combination of Scores

In this chapter we describe the effects of combining different machine scores to predict human ratings, by using *multiple linear regression*. We also compare predictions of academic marks based on combinations of machine score with predictions based on combinations of human ratings.

The regression models were trained and implemented using WEKA, a data mining software package developed at the The University of Waikato [21]. This software is freely available online from [www.cs.waikato.ac.nz/ml/weka/](http://www.cs.waikato.ac.nz/ml/weka/).

We begin the chapter with a short introduction to linear regression and its application to machine scores, human ratings and academic marks. Next, we present results of the regression experiments, focusing on the improvements in prediction accuracy achieved by different combinations of predictors.

### 8.1 Linear Regression

Simple linear regression (SLR) aims to model the conditional expected value of a target variable,  $Y$ , as a linear function of a predictor variable,  $X$ :

$$E(Y|X = x) = a_1x + b$$

We refer to estimates of parameters  $a_1$  and  $b$  as  $\hat{a}_1$  and  $\hat{b}$  respectively. One way of estimating these parameters is the *ordinary least squares* (OLS) method, which aims to minimize the *residual sum of squares* (RSS) [22].

Figure 8.1 shows an example of SLR for two hypothetical variables,  $X$  and  $Y$ . The shaded circles indicate the  $(x_i, y_i)$  data-points, and the dashed line is the line defined by  $\hat{y} = \hat{a}_1x + \hat{b}$ . The *residuals* are the differences between the true values of  $y_i$  and their estimated counterparts  $\hat{y}_i$ . In Figure 8.1, these values correspond to the signed lengths of

the vertical dotted lines. The RSS is then defined as:

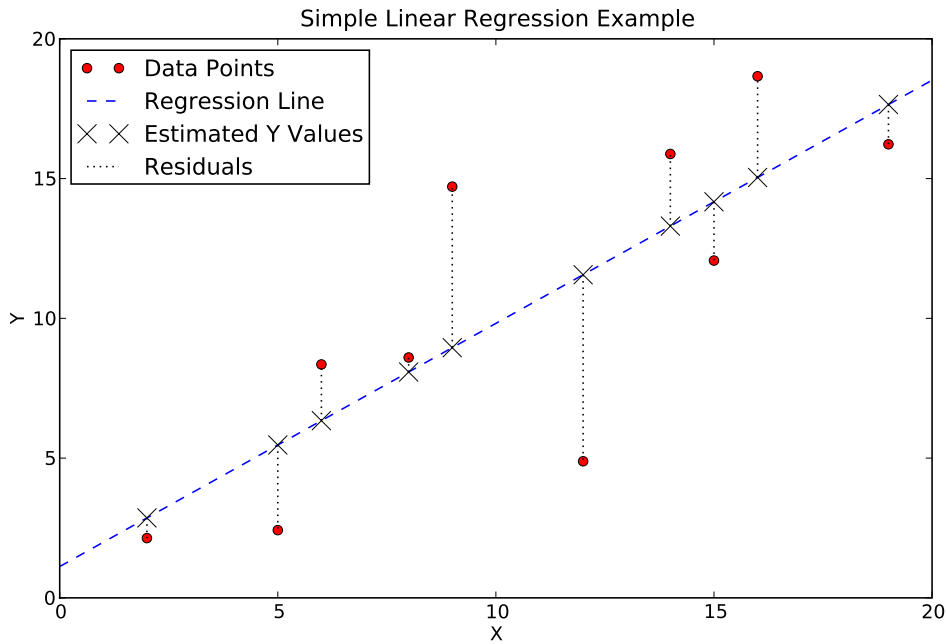
$$\begin{aligned} RSS &= \sum_{i=1}^n [y_i - \hat{y}_i]^2 \\ &= \sum_{i=1}^n [y_i - (\hat{a}_1 x_i + \hat{b})]^2 \end{aligned}$$

The values of  $\hat{a}_1$  and  $\hat{b}$  that minimize the RSS are a function of the statistical properties of the target and predictor variables, and can be shown to be [22]:

$$\hat{a}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b} = \bar{y} - \hat{a}_1 \bar{x}$$

where  $S_{xy} = \sum (x_i - \bar{x})y_i$  and  $S_{xx} = \sum (x_i - \bar{x})x_i$ . The means of the  $x$  and  $y$  values are given by  $\bar{x}$  and  $\bar{y}$  respectively.



**Figure 8.1:** *Hypothetical example of simple linear regression.*

Multiple linear regression (MLR) extends the concept of SLR to allow many predictor variables,  $X_1, X_2, \dots, X_p$ , for a single target variable  $Y$ :

$$E(Y|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = a_1 x_1 + a_2 x_2 + \dots + a_p x_p + b$$

The regression line shown in Figure 8.1 for SLR now generalizes to a  $p$ -dimensional plane in a  $(p + 1)$ -dimensional space, where  $p$  is the number of predictor variables. The RSS and the corresponding minimizing values of the  $\hat{a}_i$  and  $\hat{b}$  parameters can be calculated by means



of matrix algebra. For  $n$  data points and  $p$  predictor variables, we define the vector  $\mathbf{Y}$  and the matrix  $\mathbf{X}$ , as well as a vector of estimated regression parameters,  $\hat{\mathbf{A}}$ :

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \hat{\mathbf{A}} = \begin{pmatrix} \hat{b} \\ \hat{a}_1 \\ \vdots \\ \hat{a}_p \end{pmatrix}$$

The OLS estimate of  $\hat{\mathbf{A}}$  is then given by [22]:

$$\hat{\mathbf{A}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

## 8.2 Application of MLR to Scores, Ratings and Marks

MLR was used in our experiments for nine different configurations of target and predictor variables. Table 8.1 shows the targets and predictor categories for each of these configurations. Configurations 1-5 and 8-9 use human ratings or academic marks as targets and machine scores as predictors. Configurations 6-7 use academic marks as target variables and human ratings as predictors. The number of predictor variables is shown in parenthesis after each predictor set.

Config.	Target		Predictors
1	Reading Task Human Ratings	<i>Hesitation</i>	Reading Task Machine <b>Scores</b> (8)
2		<i>Pronunciation</i>	
3		<i>Intonation</i>	
4	Repeating Task Human Ratings	<i>Success</i>	Repeating Task Machine <b>Scores</b> (14)
5		<i>Accuracy</i>	
6	Academic Marks	<i>Oral Mark</i>	Reading and Repeating Task Human <b>Ratings</b> (5)
7		<i>Progress Mark</i>	
8	Academic Marks	<i>Oral Mark</i>	Reading and Repeating Task Machine <b>Scores</b> (22, 14 after trimming)
9		<i>Progress Mark</i>	

**Table 8.1:** *Different configurations of target and predictor variables for MLR.*

Table 8.2 lists the variables in each predictor and target category. Scores and ratings were averaged for each student, as described in Section 3.3.1. Scores which were calculated for both the reading and repeating tasks were separated into two student level averages, one for each task. These are collectively called the *reading task scores* and the *repeating task scores*. The reading task scores consist of the posterior log-likelihood scores presented in Chapter 5 and the segmentation based scores presented in Chapter 6. The repeating task

scores are the same as those for the reading task, with the addition of the accuracy scores presented in Chapter 7. Table 8.3 provides a short definition of each machine score and a reference to the section describing that score.

Reading Task Machine Scores	Repeating Task Machine Scores	Reading Task Human Ratings	Repeating Task Human Ratings	Academic Marks
$GOP_{All}$	$GOP_{All}$	<i>Hesitation</i>	<i>Success</i>	<i>Oral</i>
$GOP_{Speech}$	$GOP_{Speech}$	<i>Pronunciation</i>	<i>Accuracy</i>	<i>Progress</i>
$GOP_{Context}$	$GOP_{Context}$	<i>Intonation</i>		
$GOP_{WordLvl}$	$GOP_{WordLvl}$			
<i>ROS</i>	<i>ROS</i>			
<i>ART</i>	<i>ART</i>			
<i>PTR</i>	<i>PTR</i>			
<i>SDS</i>	<i>SDS</i>			
	<i>Acc<sub>H</sub>Results</i>			
	<i>Cor<sub>H</sub>Results</i>			
	<i>Cor<sub>Weighted-Equal</sub></i>			
	<i>Cor<sub>Weighted-Linear</sub></i>			
	<i>Cor<sub>Weighted-Quad</sub></i>			
	<i>Cor<sub>Weighted-Log</sub></i>			

**Table 8.2:** *Categories of machine scores, human ratings and academic marks.*

For each configuration, MLR models were trained for every possible combination of predictor variables. This allowed us to identify which combinations of predictors lead to the best performance, as well as combinations which have comparable success but require fewer predictors.

For configurations 1-7, all possible combinations of predictor variables were considered. However, this was not computationally feasible for configurations 8 and 9, due to the much larger number of predictors.

To address this problem, the predictor set was trimmed by removing strongly correlated candidates. For any two predictors with a correlation of more than 0.90, the predictor with the lowest correlation with the target variable was removed. This process eliminated 8 of the 22 available predictors. The remaining 14 predictors are listed in Table 8.4.

Score Name	Description	Section
$GOP_{All}$	Posterior log-likelihood score based on all phones.	5.1
$GOP_{Speech}$	Posterior log-likelihood score based on speech phones.	5.1
$GOP_{Context}$	Posterior log-likelihood score that ignores speech phones in the context of non-speech phones.	5.1
$GOP_{WordLvl}$	Posterior log-likelihood score normalised on word level.	5.1
$ROS$	Number of speech phones, normalised by the total duration of the utterance.	6.1.1
$ART$	Number of speech phones, normalised by the duration of the utterance excluding silence and non-speech events.	6.1.2
$PTR$	Duration of speech in utterance, normalised by the total duration of the utterance.	6.1.3
$SDS$	Score based on the comparison of phone durations with probability distribution of native phone durations.	6.1.4
$Acc_{HResults}$	HTK accuracy measure defined as: $(Hits - Insertions) / N_{Words} \times 100\%$	7.1.1
$Cor_{HResults}$	Percentage of words correctly repeated: $Hits / N_{Words} \times 100\%$	7.1.2
$Cor_{Weighted-Equal}$	$Cor_{HResults}$ with words grouped and weighted according to semantic importance. All weights are equal.	7.1.3
$Cor_{Weighted-Linear}$	$Cor_{HResults}$ with words grouped and weighted according to semantic importance. Weights are linearly distributed.	7.1.3
$Cor_{Weighted-Quad}$	$Cor_{HResults}$ with words grouped and weighted according to semantic importance. Weights are quadratically distributed.	7.1.3
$Cor_{Weighted-Log}$	$Cor_{HResults}$ with words grouped and weighted according to semantic importance. Weights are logarithmically distributed.	7.1.3

**Table 8.3:** Descriptions of machine scores listed in Table 8.2.

<b>Reading and Repeating Task Machine Scores After Trimming</b>
$GOP_{Context}$ (Reading Task)
$GOP_{WordLvl}$ (Reading Task)
$ROS$ (Reading Task)
$PTR$ (Reading Task)
$SDS$ (Reading Task)
$GOP_{Context}$ (Repeating Task)
$GOP_{WordLvl}$ (Repeating Task)
$ROS$ (Repeating Task)
$ART$ (Repeating Task)
$PTR$ (Repeating Task)
$SDS$ (Repeating Task)
$Acc_{HResults}$ (Repeating Task)
$Cor_{HResults}$ (Repeating Task)
$Cor_{Weighted-Linear}$ (Repeating Task)

**Table 8.4:** Predictor set consisting of machine scores for the reading and repeating tasks after trimming strongly correlated scores.

## 8.3 Evaluation

Due to the relatively small size of our corpus, leave-one-out cross validation was used to evaluate each combination of predictors.

For  $N$  speakers, leave-one-out cross validation employs  $N$  different regression models. Each model is trained on  $N - 1$  speakers and used to predict the target associated with the  $N_{th}$  speaker. A total of  $N$  iterations are considered, leaving each speaker out in turn. This leads to a set of  $N$  predicted target values, each estimated using a separately trained model. In the scenario presented here,  $N = 90$ , corresponding to the 90 students in the test set. The composition of the test set is described in Section 3.1.2.

The ability of each predictor combination to accurately estimate the target variable was evaluated by calculating the correlation between the actual target values and the predicted values.

As before, Spearman’s rank correlation coefficients were used. This allowed us to compare the correlation values determined here with those obtained for individual machine scores in Chapters 5, 6 and 7.

Tables 8.5 to 8.13 present the resulting correlation coefficients obtained with the various MLR combinations. In each case, the highest correlation with the target variables achieved using a single predictor variable is used as a *baseline*. The relative improvement achieved by

combining this predictor with others is then indicated as a percentage.

Predictions were calculated for all possible combinations of predictor variables. For each target and predictor configuration, the most successful combination was identified. We present the most effective single predictor, followed by increasingly larger combinations that were found to improve performance.

### 8.3.1 *Hesitation*

Table 8.5 shows the performance of MLR predictions for *Hesitation* ratings. The available predictors were the 8 different machine scores calculated for the reading task.

Score Name	Combinations		
<i>PTR</i>	✓	✓	✓
<i>ROS</i>		✓	✓
$GOP_{Speech}$			✓
$GOP_{WordLvl}$			✓
<b>Number of Predictors:</b>	1	2	4
<b>Correlation:</b>	0.63	0.68	0.69
<b>Improvement:</b>	<i>Baseline</i>	8%	10%

**Table 8.5:** Results for MLR predictions of *Hesitation* ratings based on reading task machine scores.

The best single predictor was the *Phonation/Time Ratio* (*PTR*). Adding *Rate of Speech* (*ROS*) to the MLR model improved the correlation by 8%. No combination consisting three scores resulted in a stronger correlation than the grouping of *PTR* and *ROS*. Adding both the *Goodness of Pronunciation* scores  $GOP_{Speech}$  and  $GOP_{WordLvl}$  to the combination lead to the best overall correlation observed with *Hesitation* ratings, representing a total relative improvement of 10% over the baseline.

### 8.3.2 *Pronunciation*

The results for MLR predictions of *Pronunciation* ratings are shown in Table 8.6. MLR models were trained on all possible combinations of the 8 machine scores calculated for the reading task.

The best performing single predictor was *Articulation Rate* (*ART*). Combining *Articulation Rate* with the *Segment Duration Score* (*SDS*) improved the correlation of predicted ratings with actual ratings by 13%. This was also the best performance of any combination of predictors for this target. The contribution made by *SDS* is surprising, considering that the correlation between *SDS* and *Pronunciation* ratings is only  $-0.18$ , as shown in Table 6.4.

Score Name	Combinations	
<i>ART</i>	✓	✓
<i>SDS</i>		✓
<b>Number of Predictors:</b>	1	2
<b>Correlation:</b>	0.47	0.53
<b>Improvement:</b>	<i>Baseline</i>	13%

**Table 8.6:** Results for MLR predictions of Pronunciation ratings based on reading task machine scores.

### 8.3.3 Intonation

Table 8.7 shows the performance of MLR with *Intonation* ratings as target and the 8 different machine scores used for the reading task as predictors.

Score Name	Combinations		
<i>ROS</i>	✓	✓	✓
<i>GOP<sub>Context</sub></i>		✓	✓
<i>GOP<sub>WordLvl</sub></i>			✓
<b>Number of Predictors:</b>	1	2	3
<b>Correlation:</b>	0.48	0.49	0.52
<b>Improvement:</b>	<i>Baseline</i>	2%	8%

**Table 8.7:** Results for MLR predictions of Intonation ratings based on reading task machine scores.

The best correlation achieved by a single predictor was that of *Rate of Speech (ROS)*. Combining this predictor with *GOP<sub>Context</sub>* improved the correlation between predicted values and target values by 2%. The addition of *GOP<sub>WordLvl</sub>* resulted in the strongest overall correlation, with a total relative improvement of 8% over the baseline.

### 8.3.4 Success

Table 8.8 shows the results for MLR predictions of *Success* ratings. The 14 machine scores calculated for the repeating task were used as predictors.

The best predictions of the target values using a single predictor were based on the *HResults Correct (Cor<sub>HResults</sub>)* scores. Combining these scores with *Rate of Speech (ROS)* lead to a 9% increase in correlation. The addition of further scores to the combination resulted in small additional improvements in the correlation. The best combination consisted of 8 machine scores, and lead to a performance that was 16% better than *Cor<sub>HResults</sub>* alone.

Score Name	Combinations						
$Cor_{HResults}$	✓	✓	✓	✓	✓	✓	✓
$ROS$		✓	✓	✓	✓	✓	✓
$Cor_{Weighted-Quad}$			✓	✓	✓	✓	✓
$Cor_{Weighted-Log}$				✓	✓	✓	✓
$Cor_{Weighted-Equal}$					✓	✓	✓
$SDS$						✓	✓
$GOP_{All}$							✓
$GOP_{Speech}$							✓
<b>Number of Predictors:</b>	1	2	3	4	5	6	8
<b>Correlation:</b>	0.75	0.82	0.83	0.84	0.85	0.86	0.87
<b>Improvement:</b>	<i>Baseline</i>	9%	11%	12%	13%	15%	16%

**Table 8.8:** Results for MLR predictions of Success ratings based on repeating task machine scores.

### 8.3.5 Accuracy

Table 8.9 shows the performance of MLR predictions of Accuracy ratings. The 14 machine scores calculated for the repeating task were used as predictors.

Score Name	Combinations			
$Cor_{HResults}$	✓	✓	✓	✓
$ROS$		✓	✓	✓
$Cor_{Weighted-Equal}$			✓	✓
$Cor_{Weighted-Linear}$			✓	✓
$Cor_{Weighted-Log}$			✓	✓
$GOP_{All}$				✓
$GOP_{Speech}$				✓
<b>Number of Predictors:</b>	1	2	5	7
<b>Correlation:</b>	0.84	0.87	0.88	0.90
<b>Improvement:</b>	<i>Baseline</i>	4%	5%	7%

**Table 8.9:** Results for MLR predictions of Accuracy ratings based on repeating task machine scores.

As was the case with the prediction of Success ratings, *HResults Correct* ( $Cor_{HResults}$ ) was the best single predictor. The addition of *Rate of Speech* ( $ROS$ ) improved the correlation between predictions and the actual ratings by 4%. The inclusion of further scores in the combination resulted in additional small increases in the correlation. The best performing combination showed a 7% relative improvement over the baseline using a combination of 7 machine scores.

### 8.3.6 Oral Mark

Two MLR configurations were applied to the prediction of the test population’s academic oral marks. The first configuration used the 5 human rating scales as predictors. The results of this configuration are shown in Table 8.10. The best single predictor was the *Accuracy* rating. Combining this rating with those for *Pronunciation* and *Intonation* resulted in an improvement of 8%. However, this correlation, 0.42, is still relatively low.

Score Name	Combinations		
<i>Accuracy</i>	✓	✓	✓
<i>Pronunciation</i>		✓	✓
<i>Intonation</i>			✓
<b>Number of Predictors:</b>	1	2	3
<b>Correlation:</b>	0.39	0.40	0.42
<b>Improvement:</b>	<i>Baseline</i>	3%	8%

**Table 8.10:** Results for MLR predictions of oral marks based on human proficiency ratings.

The second configuration used machine scores from both the reading and repeating tasks as predictors. This set of 22 scores was trimmed to a set of 14, as described in Section 8.2. The best single predictor was the *Segment Duration Score (SDS)*, based on the repeating task, as indicated in Table 8.11. The addition of *Rate of Speech (ROS)* and *HResults Accuracy (Acc<sub>HResults</sub>)*, both calculated for the repeating task, lead to a 4% increase in correlation between predicted values and target values. This correlation of 0.57 is higher than that obtained by using human ratings as predictors, as presented in Table 8.10.

Score Name	Combinations		
<i>SDS</i> (Repeating Task)	✓	✓	✓
<i>ROS</i> (Repeating Task)		✓	✓
<i>Acc<sub>HResults</sub></i> (Repeating Task)			✓
<b>Number of Predictors:</b>	1	2	3
<b>Correlation:</b>	0.55	0.56	0.57
<b>Improvement:</b>	<i>Baseline</i>	2%	4%

**Table 8.11:** Results for MLR predictions of oral marks based on machine scores.



### 8.3.7 Progress Mark

As with the oral marks, we investigated two configurations for predicting the students' academic progress marks. The first configuration used the 5 human rating scales as predictors. As shown in Table 8.12, the *Pronunciation* ratings were the most effective single predictor. Combining these ratings with those for *Success* improved the correlation by 22%. However, this correlation is still very weak.

Score Name	Combinations	
<i>Pronunciation</i>	✓	✓
<i>Success</i>		✓
<b>Number of Predictors:</b>	1	2
<b>Correlation:</b>	0.18	0.22
<b>Improvement:</b>	<i>Baseline</i>	22%

**Table 8.12:** Results for MLR predictions of progress marks based on human proficiency ratings.

The second configuration used machine scores as predictors of the progress marks. Scores for both the reading and repeating tasks were used, and the size of this set was reduced in the same manner as with the prediction of oral marks. The results for this MLR using this configuration are shown in Table 8.13.

Score Name	Combinations		
$GOP_{Context}$ (Repeating Task)	✓	✓	✓
$Acc_{HResults}$ (Repeating Task)		✓	✓
$ROS$ (Reading Task)			✓
<b>Number of Predictors:</b>	1	2	3
<b>Correlation:</b>	0.32	0.35	0.36
<b>Improvement:</b>	<i>Baseline</i>	9%	13%

**Table 8.13:** Results for MLR predictions of progress marks based on machine scores.

The best single predictor variable was the  $GOP_{Context}$  score calculated for the repeating task. The addition of the *HResults Accuracy* ( $Acc_{HResults}$ ) score for the repeating task improved the correlation between predicted values and target values by 9%. The best performance, representing a relative improvement of 13% over the baseline, was achieved by the addition of *Rate of Speech* ( $ROS$ ), calculated for the reading task, to the combination. However, the resulting correlation of 0.36 is still relatively small.

## 8.4 Summary and Conclusions

This chapter has investigated the use of combinations of machine scores for the estimation of human ratings and of academic marks using linear regression. The performance of various combinations of scores was evaluated and compared using the Spearman correlation between the predicted and the actual target values.

For human ratings, it was shown that the use of score combinations of scores results in better predictions than the use of individual scores. The greatest improvement was achieved by the first additional score, which provided an average relative increase in correlation of 7.2%. The average maximum improvement in correlation was 10.8%, and was achieved using combinations of between 2 and 8 different scores as predictors.

It was found that predictions of the reading task ratings relied mostly on segmentation based scores, while predictions of the repeating task ratings relied on segmentation based scores as well as scores based on repeat accuracy. For both the repeating task rating scales, the best single predictor was *HResults Correct*, while the addition of *Rate of Speech* resulted in the greatest improvement of correlation. The high correlations of 0.87 and 0.90 obtained for predictions of *Success* and *Accuracy* ratings respectively seem to confirm the feasibility of automated rating of a repeating task using the algorithms described in this research.

Posterior log-likelihood scores contributed little to the prediction of human ratings, except for the prediction of *Intonation* ratings.

For the prediction of academic marks, combinations of ratings and scores were considered separately. Although none of the resulting correlations between predicted values and target values were very high, it was clear that predictions based on combinations of machine scores were more accurate than those based on combinations of human ratings.

The academic oral mark is based on the lecturer's assessment of a number of oral exercises performed during the course, including prepared presentations and role-playing situations. It can be argued that this mark reflects an assessment of different aspects of proficiency than those measured by the automated test. The composition of the academic progress mark also includes assessments of written tasks. It is therefore not surprising that predicting these academic marks using scores and ratings based on the automated test is less successful than some of the other configurations investigated here. Taking this into consideration, the 0.57 correlation between oral marks and predictions based on *SDS*, *ROS* and *Acc<sub>HResults</sub>* is still informative. All of these predictors were derived from repeated utterances, stressing the importance of such a task in an automated oral test. Furthermore, while the *SDS* scores were poorly correlated with human ratings for the test utterances, they do show potential as a predictor of oral proficiency beyond the scope of the oral test.

Linear regression is one of many methods that can be used to combine machine scores in order to predict proficiency ratings. For example, Franco et al. found non-linear approaches such as the use of artificial neural networks, distribution estimation and regression trees to be slightly more effective than linear regression [10]. However, we mitigate the possible negative effect of non-linearities on our correlation values by using Spearman's rank correlation rather

than Pearson's correlation coefficient. The correlation values in this chapter should not be seen as the optimal performance of such a system, but rather as an indication of the potential of certain machine scoring algorithms to contribute to the accurate prediction of human ratings.

# Chapter 9

## Summary and Conclusions

This chapter summarises the results and conclusions presented in this thesis, and makes recommendations for future work.

### 9.1 Human Ratings

The experiments in this thesis were conducted on a corpus of recorded responses to an automated oral proficiency test. We focussed on two tasks from this test, namely the reading and repeating tasks. The students responses for these two tasks were rated for proficiency by human raters. In Chapter 3 we presented details of the rating scales used, the distribution of ratings assigned for each scale, as well as the consistency and agreement of the raters.

Ratings for the reading task were found to be concentrated in the upper region of the relevant rating scales. This poor distribution of ratings contributed to poor inter-rater agreement and weak correlations with machine scores for this task. It is concluded that the reading task must be redesigned to be more challenging in future iterations of the automated test.

The human ratings for the repeating task were better distributed and this allowed higher correlations with machine scores. Chapter 8 showed that scores for the repeating task contributed more to the prediction of the students' academic oral marks using multiple linear regression. This seems to indicate that a repeating task may be a more suitable method of evaluating the oral proficiency of advanced second language speakers than a reading task.

### 9.2 Machine Score Algorithms

Chapters 5, 6 and 7 presented a range of machine scores which use automatic speech recognition to calculate proficiency assessments. In order to evaluate the usefulness of these scoring algorithms, we considered not only their individual correlations with human ratings, but also their contributions to the prediction of human ratings when combined using linear regression, as described in Chapter 8. Table 9.1 summarises the correlations between all the different machine scores and human rating scales.

	Reading Task			Repeating Task	
	<i>Hesitation</i>	<i>Pronunciation</i>	<i>Intonation</i>	<i>Success</i>	<i>Accuracy</i>
$GOP_{All}$	0.05	0.02	0.07	0.39	0.35
$GOP_{Speech}$	0.06	0.00	0.07	0.42	0.39
$GOP_{Context}$	0.08	-0.02	0.08	0.45	0.41
$GOP_{WordLvl}$	-0.07	-0.14	-0.08	0.31	0.25
<i>Rate of Speech</i>	-0.54	-0.48	-0.49	-0.67	-0.65
<i>Articulation Rate</i>	-0.41	-0.50	-0.46	-0.60	-0.58
<i>Phonation/Time Ratio</i>	-0.64	-0.18	-0.39	-0.45	-0.44
<i>Segment Duration Score</i>	0.15	-0.18	0.00	-0.61	-0.56
<i>HResults Accuracy</i>				-0.61	-0.63
<i>HResults Correct</i>				-0.76	-0.85
<i>Weighted Correct</i>				-0.71	-0.79

**Table 9.1:** Summary of the correlations between machine scores and human ratings. Scores for the reading task are based on recognition using a finite state grammar and those for the repeating task are based on recognition using a unigram language model.

### Posterior Log-Likelihood Scoring

In Chapter 5, four variants of the posterior log-likelihood score *Goodness of Pronunciation* ( $GOP$ ) were investigated. Although many authors report this algorithm to be an effective method of evaluating the pronunciation of foreign language students, the scores did not perform well in our experiment. This discrepancy may be caused by the relatively high proficiency of second language speakers compared to that of foreign language speakers. It is possible that  $GOP$  scores are more suitable for foreign language speakers, who tend to impose the pronunciation of their mother-tongues on the target language.

Best performance was obtained for  $GOP_{Context}$ , a variant of the algorithm which calculates the utterance level  $GOP$  score without taking non-speech phones and speech phones in the left or right context of non-speech phones into account. For  $GOP$  scores which are normalised at the word level rather than the phone level ( $GOP_{WordLvl}$ ), poor correlations were observed with all rating scales. In Chapter 8 however, it led to a substantial improvement in the accuracy of multiple linear regression predictions of *Intonation* ratings, leading to a correlation of 0.52 with human ratings. This seems to indicate that where *Intonation* ratings are concerned, the word level  $GOP$  scores contain some information not provided by the other  $GOP$  scores.

### Scores Based On Segmentation

Chapter 6 presented four machine scores based on the segmentation of the utterances to be assessed. The best performance was that of *Rate of Speech (ROS)*. The *ROS* of an utterance is simple to calculate and correlated well with all rating scales for both the reading and repeating tasks. The closely related *Articulation Rate (ART)* did not lead to an improvement over *ROS*. The *Phonation/Time Ratio (PTR)* scores were strongly correlated with *Hesitation* ratings, but the *Segment Duration Score (SDS)* had little correlation with ratings for the reading task and only moderate correlation with those for the repeating task. In Chapter 8 however, *SDS* showed potential as a predictor of students' academic oral marks.

### Scores Based On Repeat Accuracy

Chapter 7 described three scores based on the accuracy of a repeated prompt. These scores were only applied to the repeating task and were found to perform reasonably well. The best correlations with the rating scales for *Success* and *Accuracy* were with the percentage of correctly repeated words ( $Cor_{HResults}$ ). An attempt was made to improve these correlations by associating a scoring weight with each word based on its semantic importance, but was unsuccessful.

## 9.3 Combination of Machine Scores

As a final step, Chapter 8 investigated the combination of different machine scores to predict human ratings, using multiple linear regression. For all rating scales, correlations were improved above that obtained when using only a single score. The strongest correlations with the reading task scales *Hesitation*, *Pronunciation* and *Intonation* were 0.69, 0.53 and 0.52 respectively. For the repeating task scales *Success* and *Accuracy*, the corresponding correlations were 0.87 and 0.90. These high correlations confirmed the feasibility of the automatic assessment of a repeating task using machine scores and linear regression. The lower correlations for the reading task may in part be due to the high average ratings and low inter-rater agreement for this task, as described in Chapter 3.

## 9.4 Recommendations for Future Research

### Automated Oral Test

The performance of machine scores on a given corpus may be affected by many factors, such as the composition of the test population, the design of the test prompts and the agreement and consistency of the raters. A comparative study using the same algorithms proposed in this thesis on a similar corpus would shed light on the robustness of the findings in this research. During 2009, a new corpus was collected at the Stellenbosch University Faculty of Education, and results based on this corpus should become available in the near future.

The high average ratings for the reading task reduced its usefulness for evaluating the performance of machine scores. For future tests, care must be taken to design a more challenging reading task. The repeating task proved to be an informative method of assessing second language students as well as the performance of machine scores. Other types of tasks may also be effective, such as the *shadowing* task which was used to assess foreign language speakers in a recent study [23]. For this task, students were required to repeat a continuous stream of dialogue as it was being heard. The shadowing task was shown to be a more effective method of assessment than a reading task used in the same study.

### Machine Score Algorithms

Certain machine scores investigated in this thesis performed better than others. For future research where it is desirable to focus on a reduced set of scores, the following scores are recommended, based on their individual correlations with human ratings as well as their contributions to successful multiple linear regression combinations:

- $GOP_{Context}$
- $GOP_{Wordlevel}$
- *Rate of Speech*
- *Phonation/Time Ratio*
- *Segment Duration Score*
- $Cor_{HResults}$

However, it may be instructive to perform a comparative study using all of the machine scores in this thesis, to assess the degree to which the findings presented here generalise to other data sets.

### Combinations of Machine Scores

In this thesis, different machine scores were combined using linear regression, and the performance of these combinations were evaluated using Spearman's rank correlation coefficient. This approach indicated the potential accuracy with which combinations of machine scores could predict human ratings. However, in a fully operational system it may be more desirable to perform *classification* rather than *regression*, as this corresponds to the project goal of classifying students according to their oral proficiency. Also, while time constraints allowed only the application of linear regression as a method of combining scores for this thesis, non-linear methods such as artificial neural networks, distribution estimation and regression trees should be investigated in future research.

# Bibliography

- [1] C. Van der Walt, F. De Wet, and T. R. Niesler, “Oral proficiency assessment: the use of automatic speech recognition systems,” *Southern African Linguistics and Applied Language Studies*, vol. 26, pp. 135–146, 2008.
- [2] F. De Wet, C. Van der Walt, and T. R. Niesler, “Automatic large-scale oral language proficiency assessment,” in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 218–221.
- [3] F. De Wet, C. Van der Walt, and T. R. Niesler, “Automatic assessment of oral language proficiency and listening comprehension,” *Speech Communication*, vol. 51, pp. 864–874, 2009.
- [4] F. De Wet, P. F. De V. Müller, C. Van der Walt, and T. R. Niesler, “Experiments in automatic assessment of oral proficiency and listening comprehension for bilingual South African speakers of English,” in *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Cape Town, South Africa, 2008, pp. 67–72.
- [5] P. F. De V. Müller, F. De Wet, C. Van der Walt, and T. R. Niesler, “Automatically assessing the oral proficiency of proficient L2 speakers,” in *Proceedings of SLATE*, Warwickshire, UK, 2009, CD-ROM.
- [6] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, pp. 95–108, 2000.
- [7] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic scoring of pronunciation quality,” *Speech Communication*, vol. 30, pp. 83–93, 2000.
- [8] C. Cucchiaroni, H. Strik, and L. Boves, “Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms,” *Speech Communication*, vol. 30, pp. 109–119, 2000.
- [9] C. Hacker, T. Cincarek, R. Gruhn, S. Steidl, E. Nöth, and H. Niemann, “Pronunciation feature extraction,” in *Proceedings of 27th DAGM Symposium*, Vienna, Austria, 2005, pp. 141–148.



- [10] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. J. Cesari, “The SRI EduSpeak<sup>TM</sup> system: Recognition and pronunciation scoring for language learning,” in *Proceedings of InSTILL 2000*, Dundee, Scotland, 2000, pp. 123–128.
- [11] T. Cincarek, R. Gruhn, C. Hacker, E. Nöth, and S. Nakamura, “Automatic pronunciation scoring of words and sentences independent from the non-native’s first language,” *Computer Speech and Language*, vol. 23, pp. 65–88, 2009.
- [12] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology,” *Journal of the Acoustical Society of America*, vol. 107, pp. 989–999, 2000.
- [13] C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners’ fluency: comparisons between read and spontaneous speech,” *Journal of the Acoustical Society of America*, vol. 111, pp. 2862–2873, 2002.
- [14] *STATISTICA 8.0*. StatSoft Incorporated, 2008, [www.statsoft.com](http://www.statsoft.com).
- [15] K. Zechner, D. Higgins, and X. Xi, “Speechrater: A construct-driven approach to scoring spontaneous non-native speech,” in *Proceedings of SLaTE*, Farmington, PA, USA, 2007, pp. 128–131.
- [16] D. G. Rees, *Essential statistics - Fourth Edition*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2001.
- [17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.
- [18] J. C. Roux, P. H. Louw, and T. R. Niesler, “The African Speech Technology project: An Assessment,” in *Proceedings of LREC*, Lisbon, Portugal, 2004, pp. I:93–96.
- [19] O. D. Deshmukh, S. Joshi, and A. Verma, “Automatic pronunciation evaluation and classification,” in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 1721–1724.
- [20] J. Richards and R. Schmidt, *Longman Dictionary of Language Teaching and Applied Linguistics*. London, UK: Longman, 2002.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explorations*, vol. 11, 2009, available online at [www.sigkdd.org](http://www.sigkdd.org), last accessed 2009/11/25.
- [22] S. Weisberg, *Applied linear regression*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005.

- [23] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, “Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences,” in *Proceedings of SLaTE*, Warwickshire, UK, 2009, CD-ROM.

# Appendix A

## Reading Task Prompts

This appendix contains the eleven sentences of the reading task. Each student taking the test was asked to read a random selection of six of these sentences from his or her test sheet.

1. Many schools in South Africa require support to create a positive learning environment.
2. However, appropriate resources are expensive and must be maintained properly.
3. School governing boards struggle to make ends meet.
4. It is up to the government to ensure a fair allocation of funds.
5. During the staff meeting teachers discussed the new grade eight intake.
6. It emerged that there has been a twenty percent drop in the initial enrolments.
7. Could the school be losing its reputation as a major role player in the area?
8. The principal will have to reassure all the parents and teachers.
9. At the regional workshop for Western Cape teachers we discussed the framework for the new senior certificate.
10. Many participants asked if this was the best way forward.
11. Their minds were put at rest when the implementation plans were presented.

# Appendix B

## Repeating Task Prompts

This appendix contains the eight sentences of the repeating task. The word and phrase ranks used for the *Weighted Correct* score, as discussed in Section 7.1.3, are shown underneath each sentence. The highest rank, 6, indicates the word or phrase of the most semantic importance.

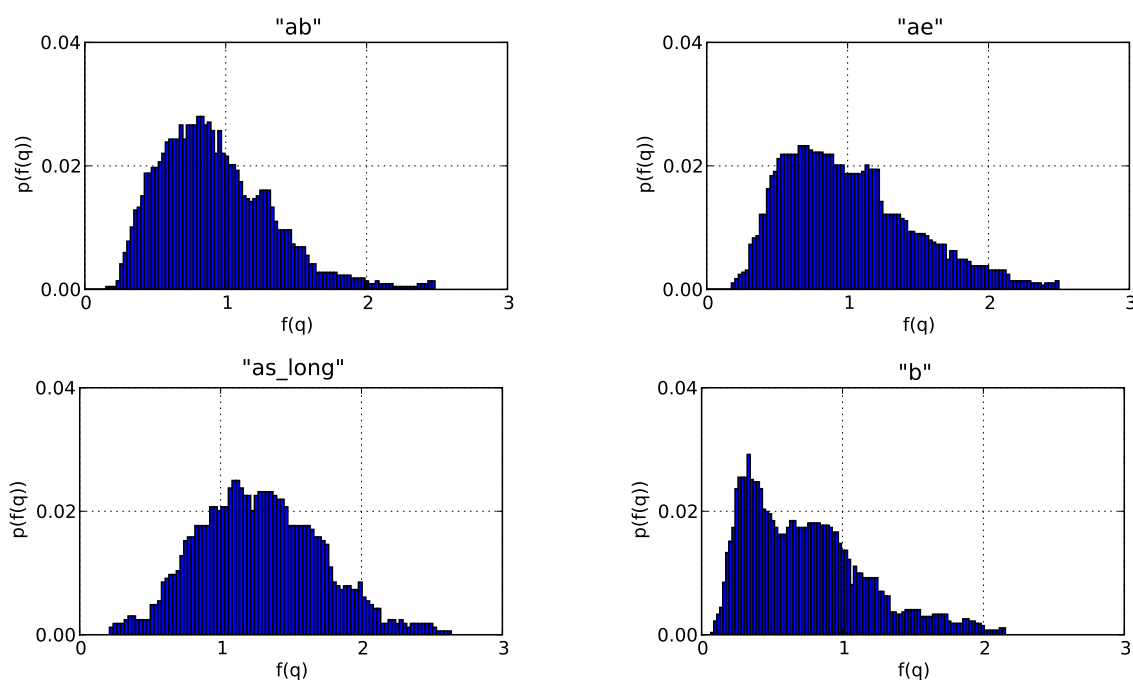
1. Student teachers do not get enough exposure to teaching practice.  
5 6 1 4 1 3 1 2
2. During visits to the schools they are seldom required to teach advanced classes.  
1 6 1 5 1 3 1 4 1 2
3. Lecturers who are out of touch with school practice have unrealistic expectations.  
6 1 5 2 1 4 1 2 3
4. It is boring to sit and watch teachers all day.  
6 5 4 1 3 2 1
5. Learners appear uninterested and there is an alarming lack of motivation and ambition.  
6 5 4 1 3 1 2 1 2 1 2
6. Could the materialistic society we live in be responsible for their attitude.  
6 1 5 1 4 1 3
7. The efficiency of a school depends to a large extent on the capabilities of the principal.  
1 6 1 5 1 2 1 4 1 3
8. How parents interests and hopes are accommodated is crucial to the success of a school.  
1 6 2 1 2 5 5 4 1 3 1 2

# Appendix C

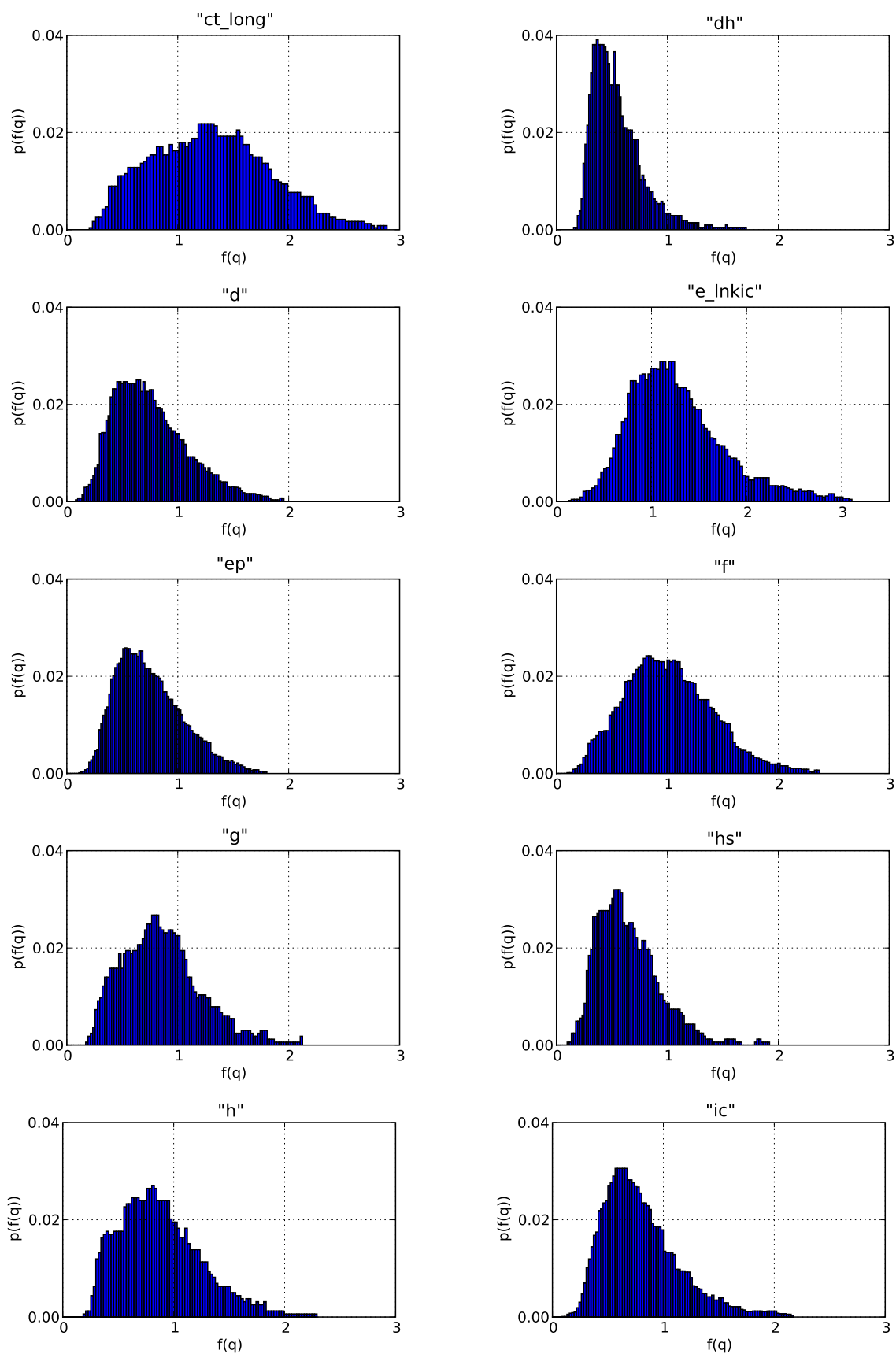
## Distributions of Monophone Durations

This appendix presents plots of the discrete probability distributions of normalised durations of the 34 monophones used to calculate the Segment Duration Score. Details of how these distributions were calculated are presented in Section 6.1.4.

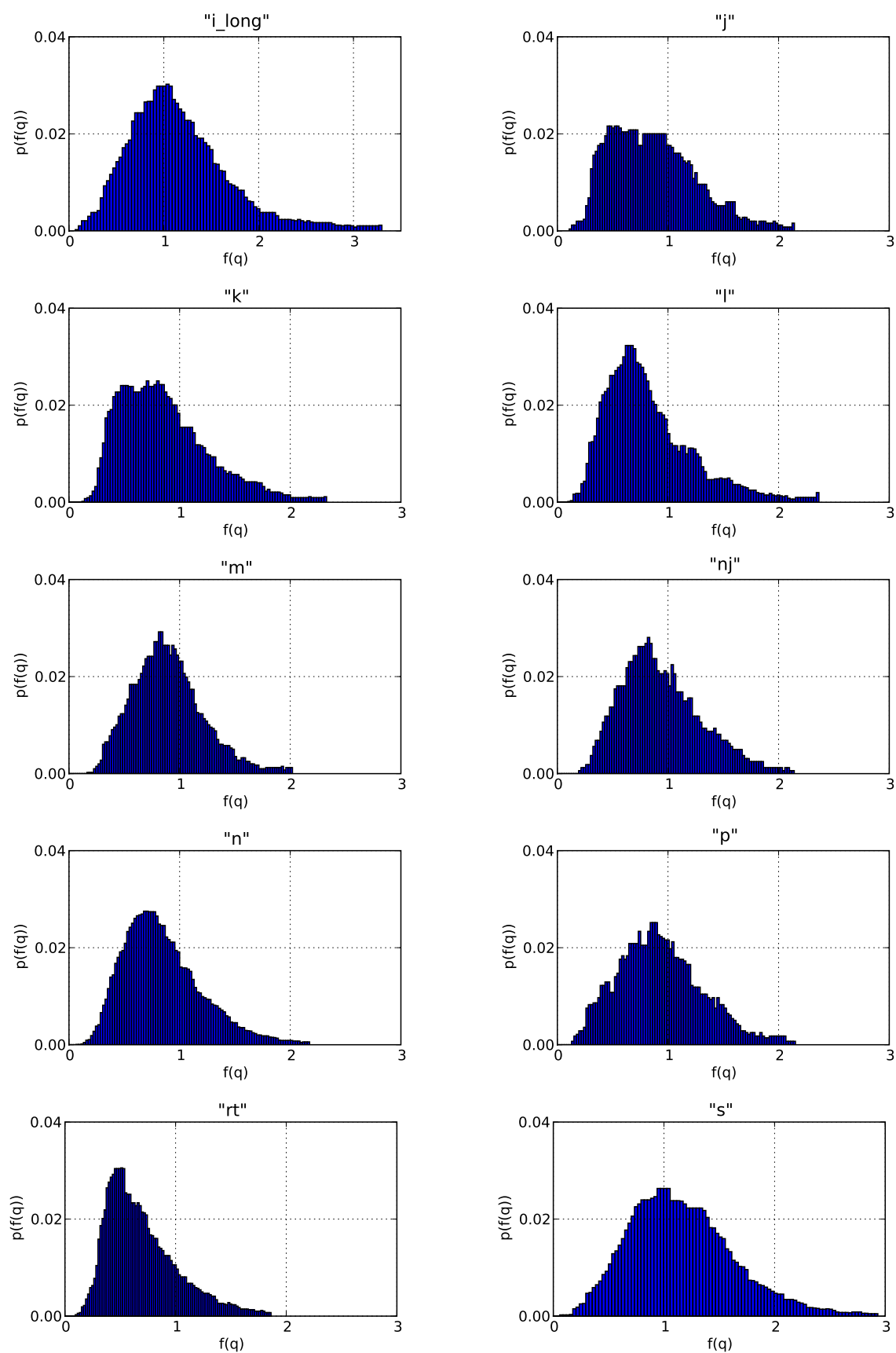
In Figures C.1 to C.4, the x-axis value,  $f(q)$ , refers to the duration of a specific instance of the relevant monophone,  $q$ , normalised by the Rate of Speech of the utterance in which that monophone occurs.



**Figure C.1:** Discrete probability distributions of the normalised duration,  $f(q)$ , of monophones.



**Figure C.2:** Discrete probability distributions of the normalised duration,  $f(q)$ , of monophones.



**Figure C.3:** Discrete probability distributions of the normalised duration,  $f(q)$ , of monophones.

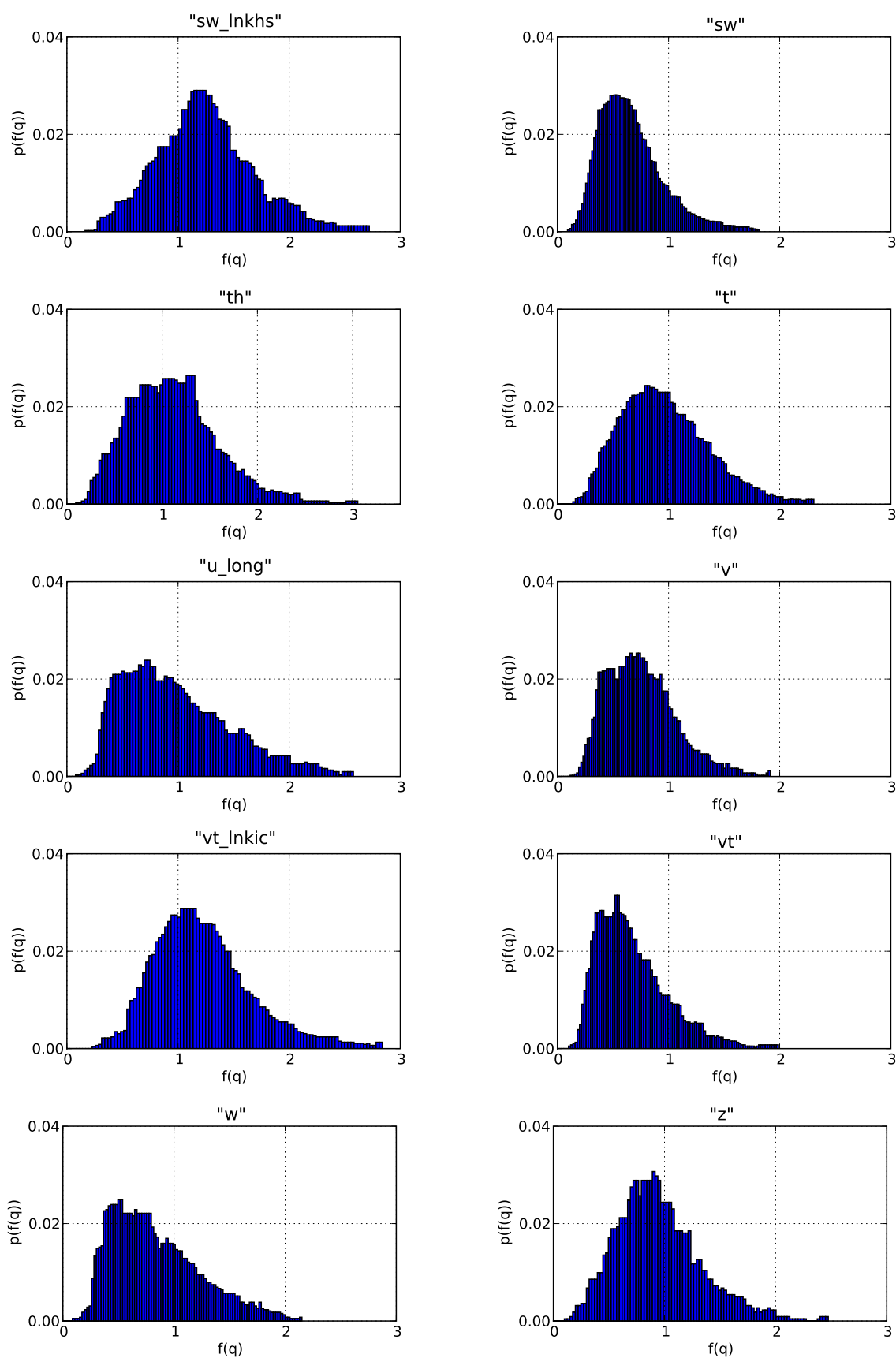


Figure C.4: Discrete probability distributions of the normalised duration,  $f(q)$ , of monophones.



# Appendix D

## Inter-Score Correlations

This appendix shows the inter-score correlation, calculated using Spearman’s rank correlation coefficient. Table D.1 shows the inter-score correlations for the reading task, and Table D.2 shows the inter-score correlations for the repeating task.

	$GOP_{All}$	$GOP_{Speech}$	$GOP_{Context}$	$GOP_{WordLvl}$	$ROS$	$ART$	$PTR$	$SDS$
$GOP_{All}$	1.00	0.99	0.97	0.92	0.03	0.11	-0.23	-0.21
$GOP_{Speech}$	0.99	1.00	0.99	0.92	0.05	0.14	-0.24	-0.21
$GOP_{Context}$	0.97	0.99	1.00	0.90	0.05	0.15	-0.25	-0.19
$GOP_{WordLvl}$	0.92	0.92	0.90	1.00	0.17	0.25	-0.16	-0.11
$ROS$	0.03	0.05	0.05	0.17	1.00	0.96	0.58	-0.12
$ART$	0.11	0.14	0.15	0.25	0.96	1.00	0.36	-0.03
$PTR$	-0.23	-0.24	-0.25	-0.16	0.58	0.36	1.00	-0.27
$SDS$	-0.21	-0.21	-0.19	-0.11	-0.12	-0.03	-0.27	1.00

**Table D.1:** *Inter-score correlations for the reading task.*

	$GOP_{All}$	$GOP_{Speech}$	$GOP_{Context}$	$GOP_{WordLvl}$	$ROS$	$ART$	$PTR$	$SDS$	$AccHResults$	$Cor^HResults$	$Cor^{Weighted-Equal}$	$Cor^{Weighted-Linear}$	$Cor^{Weighted-Quad}$	$Cor^{Weighted-Log}$
$GOP_{All}$	1.00	1.00	0.95	0.82	-0.38	-0.38	-0.21	-0.44	-0.42	-0.39	-0.38	-0.29	-0.19	-0.33
$GOP_{Speech}$	1.00	1.00	0.96	0.83	-0.41	-0.39	-0.26	-0.45	-0.45	-0.41	-0.39	-0.30	-0.21	-0.34
$GOP_{Context}$	0.95	0.96	1.00	0.75	-0.46	-0.43	-0.28	-0.49	-0.47	-0.39	-0.30	-0.22	-0.06	-0.34
$GOP_{WordLvl}$	0.82	0.83	0.75	1.00	-0.25	-0.20	-0.25	-0.28	-0.35	-0.31	-0.15	-0.06	-0.20	-0.20
$ROS$	-0.38	-0.41	-0.46	-0.25	1.00	0.85	0.68	0.56	0.77	0.56	0.51	0.47	0.54	0.54
$ART$	-0.38	-0.39	-0.43	-0.20	0.85	1.00	0.24	0.65	0.67	0.50	0.50	0.50	0.45	0.52
$PTR$	-0.21	-0.26	-0.28	-0.25	0.68	0.24	1.00	0.19	0.50	0.37	0.26	0.24	0.24	0.31
$SDS$	-0.44	-0.45	-0.49	-0.28	0.56	0.65	0.19	1.00	0.54	0.51	0.33	0.33	0.48	0.48
$AccHResults$	-0.42	-0.45	-0.47	-0.35	0.77	0.67	0.50	0.54	1.00	0.66	0.62	0.60	0.52	0.62
$Cor^HResults$	-0.39	-0.41	-0.41	-0.31	0.56	0.50	0.37	0.51	0.66	1.00	0.92	0.87	0.72	0.91
$Cor^{Weighted-Equal}$	-0.38	-0.39	-0.39	-0.26	0.54	0.50	0.32	0.50	0.62	0.92	1.00	0.91	0.76	0.97
$Cor^{Weighted-Linear}$	-0.29	-0.30	-0.30	-0.15	0.51	0.50	0.26	0.43	0.60	0.87	0.91	1.00	0.94	0.98
$Cor^{Weighted-Quad}$	-0.19	-0.21	-0.22	-0.06	0.47	0.45	0.24	0.33	0.52	0.72	0.76	0.94	1.00	0.87
$Cor^{Weighted-Log}$	-0.33	-0.34	-0.34	-0.20	0.54	0.52	0.31	0.48	0.62	0.91	0.97	0.98	0.87	1.00

Table D.2: Inter-score correlations for the repeating task.

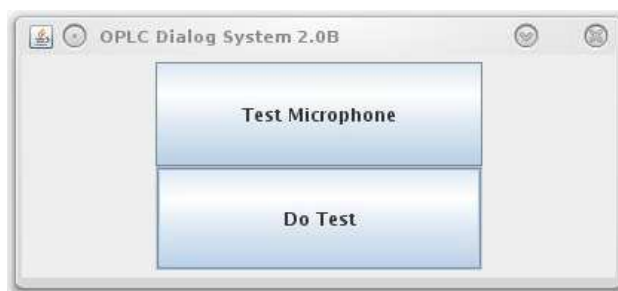
# Appendix E

## Automated Oral Test Software

This appendix describes software designed for the administering of automated oral tests. During 2008 an attempt was made to obtain a new corpus of recorded responses to an automated oral test. A telephone dialogue system was used to administer the test, but technical difficulties resulted in the loss of the majority of the test subjects' responses.

The author subsequently designed and implemented a Java application for the administering of future tests. Java was chosen as programming language because it allows the software to execute unchanged on both Windows and UNIX systems. The software was used successfully during 2009 to record a new corpus of data. Results for this new corpus are not included in this thesis, but should become available in related future research.

The software is intended to be used by students on personal computers. The test prompts are played and the user's response to each prompt is recorded. The startup window, shown in Figure E.1, allows the user to choose either of two dialogues. The "Test Microphone" dialogue records the user's name and plays the recording back, allowing the user to adjust the computer's microphone settings if necessary. The "Do Test" option starts the oral proficiency test. The user's responses are stored in a folder named according to the user's student number, which is entered upon starting the application.



**Figure E.1:** *Startup window of automated oral test software.*

Figure E.2 shows the window displayed during the test. While a prompt is being played, the window displays the text "LISTEN TO THE INSTRUCTIONS" on a red background. When the prompt ends, the background changes to green and the text "SPEAK NOW,

CLICK NEXT TO CONTINUE” is displayed. The user’s response is recorded until the “Next” button is clicked. The next prompt then begins to play.



**Figure E.2:** *Dialogue window of automated oral test software.*

The program reads in a text-based instruction file which indicates the order in which prompts must be played, as well as the path names for the prompts and recorded responses. The instruction file also allows the specification of random prompt selections, such as randomly selecting 3 out of 5 available prompts, playing each in turn and recording the responses. The use of an instruction file means that the test structure can be modified without recompiling the software.