



ISLE

Recognition of Learner Speech

Project: LE4-8353
Deliverable: D3.3

Version	4
Date	15.09.1999

ISLE Deliverable

Project Number	LE4-8353
Project Title	Interactive Spoken Language Education [ISLE]
Deliverable Type	RE
Distribution	P
Deliverable ID	D3.3
Expected Delivery Date	T13
Actual Delivery Date	24.08.99
Title of Deliverable	Recognition of Learner Speech
Author(s)	Rachel Morton (Entropic)

OT	RE	SP	PR	TO
Other	Report	Specification	Prototype	Tool

C	P	R
Consortium	Public	Restricted

Revision History

Version	Date	Status	Author(s)
1	28.07.99	Draft	Entropic [Rachel Morton]
2	13.08.99	Draft	Entropic [Rachel Morton]
3	24.08.99	Draft	Entropic [Rachel Morton]
4	15.09.99	Draft	Entropic [Rachel Morton, David Wood]

Summary

This paper describes principles for the design of interactive pronunciation exercises using the EC-ISLE system with highly accentuated learner speech. The paper “Pronunciation Teaching: Requirements and Solutions” outlined the exercise-types specified as desirable in previous user-studies. This paper is primarily concerned with measuring which of these exercise-types are really viable given the state-of-the art in recognition. Conclusions are drawn from a series of recognition experiments using a non-native corpus of Italian and German learners of English and Entropic’s HMM-based speech API (HAPI 2.0).

The factors examined include

1. Recognition task perplexity and its effect on recognition accuracy.
2. The effect of mother tongue and proficiency on recognition accuracy.
3. The effect of a mismatch in acoustic modelling between the target British English HMMs and highly accentuated incoming speech.
4. The effect of speaker adaptation techniques to transform the native acoustic models in the direction of the accented speech.

This report is for public distribution. The target reader is anyone interested in using automatic speech recognition for pronunciation teaching, in particular multimedia publishers and English language teaching professionals who wish to develop pronunciation teaching products and exercises using ISLE.

Contents

ISLE Deliverable.....	2
Revision History.....	2
Summary	3
Contents.....	4
1 Introduction.....	5
2 Non-native corpus summary	6
2.1. Linguistic form and content.....	6
2.2. Speakers.....	8
2.3. Recording Details	8
3 Recognition SetUp	9
3.1. Recognition Grammars.....	9
3.2. Baseline Recogniser	11
4 Low Perplexity Recognition with a British Recogniser.....	12
4.1. Baseline Recognition and Perplexity.....	12
4.2. Mother Tongue	12
4.3. Language Proficiency	13
4.4. Acoustic models	14
4.5. Vocabulary Size.....	16
4.6. Length constraints	16
4.7. Summary of optimal non-adapted results.....	16
Adaptation Experiments.....	17
4.8. MLLR.....	18
4.9. Model Merging.....	24
5 Conclusions and Recommendations	26
Appendix 1: References	28

1 Introduction

The area of automatic pronunciation teaching has received considerable attention in both commercial and research arenas recently. Pronunciation is an essential part of second language teaching as poor pronunciation can hinder intelligibility of the learner, however teaching pronunciation can be repetitive, requiring drills and one-to-one attention that is not always available, especially in large classes or if no teacher is available. Computer-aided pronunciation teaching tools are therefore attractive as they allow self-paced practice outside a classroom.

In a recent study [2] it was argued that successful second language acquisition, including production skills, requires certain conditions to be met. Learners should engage in tasks designed to maximise opportunities for interaction, and the interaction should be led by communicative goals that encourage the learner to produce target language output. Interactive spoken language exercises are therefore the ideal situation in which to practice and improve pronunciation. Two additional criteria were that learners must be made aware of the linguistic characteristics of the target language, by the use of salient input, and that they need to notice errors in their own output. These points are especially relevant for adult learners who may find it difficult to perceive phonetic category distinctions in a second language.

Many CALL products on the market now use automatic speech recognition to allow interactive spoken language exercises. ‘Subarashi’ [6] and ‘Traci Talk The mystery’ are examples that allow dialogue practice within goal-oriented exercises or games. The response grammars in these cases are simple but cleverly designed, making good use of the current recognition capabilities. Research into pronunciation assessment has moved towards text-independent algorithms that use phoneme-based recognition to assess speech at the segmental level [17] [15] [1] [12]. The ISLE project continues this trend, by developing a system that is able to provide diagnosis of phonemic errors and lexical stress errors. Feedback to the user will contain the location and type of error and where possible the reason for the error with follow-up exercises which will help make the learner aware of their errors. The diagnosis components will work in a generic way that is exercise-independent so that pronunciation exercises may be embedded in any communicatively relevant spoken language exercise regardless of the textual content.

The spoken language exercises will be designed and constructed using the ISLE applications programming interface (API). The ISLE software contains three main modules, the speech recognition module, an error localisation module and an error diagnosis module. The speech recognition module is based on Entropic’s ISLE recognition API (IHAPI) which uses standard HMM-based algorithms. The three modules work in a sequential manner, where each module narrows down the location of the errors. The recogniser selects the best from several utterance hypothesis allowed by the grammar. The localisation module locates areas of low confidence within this single hypothesis. These areas are then investigated at the segmental level by the diagnosis module. It is therefore essential that the speech recognition results are accurate, or the diagnosis could be very misleading and potentially discouraging to the learner.

The recognition accuracy can be affected by how well the learner’s speech is modelled, and also how difficult the exercises are in terms of grammar perplexity, active vocabulary size and phonetic confusion within the target vocabulary. To guarantee high recognition accuracy, the standard English recogniser may need to be adapted to model the acoustic characteristics of non-native speech and the recognition tasks may have to be constrained to limit the number of alternative sentence hypotheses.

A set of experiments were carried out to find a suitable recognition setup for ISLE and to understand how the recognition performance scales as tasks become less constrained and the speakers become less proficient. The outcome of these experiments will be used to guide the exercise design in the final ISLE demonstrator, and to provide useful guidelines for any potential ISLE developers.

2 *Non-native corpus summary*

The algorithms and solutions being developed for the ISLE project aim to be portable to new source and target language pairs, but the initial prototype system will focus on Italian and German adult intermediate learners of English. These languages were chosen because they are quite diverse in terms of their phonological systems and language roots. Italian speakers have more problems with stress and rhythm as well as phonemic confusions [3]. Germans on the other hand find English easier to learn initially as there are many similarities between English and German in terms of syntax phonology and vocabulary [15]. Adult intermediate learners were chosen as they will benefit from pronunciation teaching but will also have sufficient knowledge to complete reading exercises. Acoustic modelling data is also more widely available for adult speech.

A corpus of non-native speech was collected in order to evaluate and optimise the recogniser for the target speakers on target low-perplexity recognition tasks and also to evaluate the diagnosis algorithms.

2.1. Linguistic form and content

There are many possible exercise types that could be used in the ISLE system. The tasks that were considered during the experiments came from a user study in which English language teaching professionals were asked to identify tasks considered essential and tasks considered desirable in the ISLE system given that recognition accuracy supports them. This study specified that the prototype ISLE system will “accept spoken input in a controlled environment with the exercise types listed in Table 1. A fuller discussion of these results can be found in deliverable D1.4 “Pronunciation Teaching: Requirements and Solutions”.

Exercise Type	Description
Reading	Reading a text possibly presented as part of a simulated dialogue
Matching items	Oral combination of suitable items from several lists, such as choosing the subject, the verb, and an object to form “she drinks coffee” or “he plays violin”
Multiple choice exercises	Oral selection from a single list of different items
Minimal pairs	Pronunciation of pairs such as <i>dip</i> vs. <i>deep</i>
Answering questions	For example “what is the girl drinking” with a picture showing that “she is drinking soda”
Simple scene descriptions	Producing descriptions from restricted vocabulary such as “there is a house with a swimming pool”

Table 1. Exercise Types

Controlled exercise types were specified partly because the nature of pronunciation teaching with adults is usually quite focussed, highlighting certain differences between the target and source language at a time. Teachers rarely attempt to correct all a learner's problems at once. Non-native speech is also very difficult to recognise due to wide variation amongst learners in terms of the temporal and spectral characteristics of disfluent accented speech. By constraining the range of expected inputs and reducing the scope for word insertions and deletions, word accuracy can be increased [7].

Reading, matching items multiple choice and minimal pair practice were identified as essential exercise types whereas simple question answering and scene descriptions were identified as desirable extensions if the recognition accuracy permits these exercise types. The recording prompts were designed to cover all these exercise types in order that the recognition performance on each task could be evaluated.

In addition the recording texts were designed to cover general pronunciation problems of English learners such as weak forms, consonant clusters and stress, but also problematic phones resulting from the transfer from Italian and German mother tongues.

The corpus material is divided into adaptation data and task-oriented exercise data, both of which are composed of different *blocks*. Each block of exercise data covers different exercise types and pronunciation problems, summarised in Table 2.

Block	#Sents.	Linguistic Issue	Exercise Type	Examples
A	27	Wide vocabulary coverage (410)	Adaptation Reading	"In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred meters from the top before they had to turn back."
B	33			
C	22			
D	81	Problem phones Weak Forms Consonant clusters	Multiple choice Minimal Pair	"A cup of coffee" "A mouth" "A mouse" "A railway station"
D1	30	Problem phones	Minimal Pair	"I said bad not bed"
D2	15	Problem phones Weak Forms	Description Selection	"Beside a tree in a park" "Next to the jug on the table"
D3	6	Problem phones Weak Forms Consonant clusters	Description Selection	"She's wearing a brown wooly hat and a red scarf."
E	63	Stress Weak Forms Problem Phones Consonant clusters	Reading	"The convict expressed anger at the sentence." "The jury took two days to convict him."
F	10	Weak Forms Problem Phones	Reading Selection	"I would like chicken with fried potatoes, broccoli, peas and a glass of water."
G	11	Weak Forms Problem Phones	Reading Selection	"This year I'd like to visit Rome for a few days."

Table 2. Corpus Content

2.2. Speakers

The corpus consists of speech from 52 speakers, containing 23 Italians 23 Germans and three others (Spanish French and Chinese). The speakers were either volunteer staff from the partners in Stuttgart Hamburg and Milan, or overseas students at Leeds University.

The target ISLE proficiency level was intermediate, however some beginners, advanced and native speakers were included in the corpus so that effects of proficiency could be monitored. The speakers gave rough judgements of their own language proficiency as shown in Table 3. The corpus contains a total of 26 intermediate learners, 20 advanced learners, 3 beginners and 2 native speakers. One German speaker also claimed he was ‘native’.

L1	Sex		Age			Proficiency				Total
	M	F	18-29	30-39	40+	Beg.	Int.	Adv.	Nat.	
German	13	10	?	?	?	0	15	7	1	23
Italian	19	4	20	7	1	3	11	9	0	23
Other	1	3	3	1	0	0	0	4	0	4
Brit Eng	2	0	0	1	1	0	0	0	2	2
Total	35	17				3	26	20	3	52

Table 3. Corpus Summary

The self-diagnosed proficiency levels were used only as a preliminary guide for comparing speaker groups. Pronunciation-proficiency judgements from a language teaching expert are to be made available at a later date. These judgements will use a standard rating scale modified specifically for judging pronunciation.

2.3. Recording Details

Individual sessions were sent to five recording sites. The sessions were designed such that each speaker had a different order of presentation. To reduce order effects, the text within blocks and blocks themselves were randomised. To reduce fatigue, difficult blocks A B C and E were separated by blocks containing shorter easier utterances. Each speaker recorded a total of 247 prompts (a 27 b 33 c 22 d 81 e 63 f 10 g 11) in a single session lasting about 1.5 hours. The data was collected using an Entropic recording tool and a Knowles VR3565 headset microphone. The waveforms were sampled at 16 kHz and stored in WAV format files.

3 Recognition Set Up

The following sections describe the components of the baseline recognition system

3.1. Recognition Grammars

Finite-state recognition grammars were created for text blocks C to G, while blocks A and B were reserved as adaptation data. Five types of grammars were produced corresponding to the exercise types listed in Table 2. These are described in Table 4.

Grammar	Exercise Type	Description	Blocks
Psent	Reading	A parallel-sentence syntax in which the number of branches is equal to the number of sentences.	C D E F G
Mpair	Minimal Pair	A syntax where one or more choice points are embedded within the set carrier phrase, such as ‘I said X not Y’. The test set perplexity is governed by the number of choices at each point. Additional task complexity comes from the phonetic confusability of the choices at each point.	D01
Order	Description	A grammar which combines several phrase level sub-syntaxes within a semi-constrained phrase order. Each phrase sub-syntax is itself a parallel syntax of phrase alternatives. For example a greeting sub-syntax may consist of parallel branches for “Hello” “Good afternoon” “Hi” and “Can I help you?”.	F G
Syntax	Description	The <i>syntax</i> contains sub-syntaxes with much freer word order, where the branching factor is not limited to the number of choices.	F G
Wloop	Description	An unconstrained word-loop syntax containing all words in the test set	C

Table 4. Recognition Grammar Descriptions

A parallel sentence grammar was produced for all test blocks, giving test set perplexities between 1 to 2 depending on the length of the test sentences. More complex finite-state grammars were designed for each text block, depending on the syntactic form of the text. Some examples from Block D will be given to illustrate the range of ISLE tasks that are represented in the test sets. These use the extended Backus-Naur Form (EBNF) grammar specification.

D1 Minimal Pair Task

Part 1 of Block D contains sentences of the type “I said got not goat”, “I said bad not bed” “I said wine not white”. These can be defined by a grammar of the form

Sentence = (I said \$Choice1 not \$Choice2)
 Choice1 = (alone | bad | book | call | cheap | sleep | snow | sport | thin | through)
 Choice 2 = (gone | bed | do | shall | cheese | tin | slept | site | leave | bed | sheep)

The test set contains 30 unique utterances, although the number of valid sentences that could be parsed is 30*29. The task can be thought of as a minimal-pair task or a selection exercise.

D2 Syntax

Part D2 of Block D contains sentences with an optional question followed by a prepositional phrase. This grammar allows valid utterances such as “At home”, “Beside a tree in a park”, “In a boat on the river”, “Where is the boat? On the river”. With this exercise type the learner could give quite unconstrained scene descriptions with a limited vocabulary. The recursive prepositional phrase allows sentences of unlimited length but this gives room for word insertions and deletion errors so the task is one of the more difficult exercise types specified.

Sentence	= ([“where are they sitting” “where is the cup”] \$PrepPhrase)
PrepPhrase	= (\$Prep [\$Position] \$NounPhrase “at home”)
NounPhrase	= \$Det \$Noun \$PrepPhrase
Det	= (a the)
Noun	= (tree park river boat path pub cupboard table jug table home cup)
Prep	= (in on beside next to outside)
Position	= (the [middle edge] of)

D3 Syntax

Part 3 of Block D contains sentences of the form

Sentence	= ((what’s [he she] [he’s she’s]) wearing) \$Noun Phrase
NounPhrase	= ((A {\$Adj} \$Noun) [and \$NounPhrase])
\$Adj	= (big beige cowboy yellow flowery brown red .. corduroy)
\$Noun	= (hat jumper shirt trousers scarf jacket)

This grammar also allows sentences of unconstrained length due to the recursive adjectives and the repeated noun phrase. This could be used in a selection or description task where the learner must select from a list of alternative colours and items of clothing. For example “She’s wearing a brown wooly hat and a red scarf”, “He’s wearing a scarf and corduroy trousers”, “He’s wearing trousers”, “She’s wearing a big beige jumper and a flowery shirt”.

Block F and G use similar perplexity syntax grammars within the domain of holidays and the restaurant. The holiday grammar allows choices to be selected from the sentence subject, the destination and the time period. The Restaurant grammar allows a meal to be ordered that combines a choice of courses presented in different orders while selecting dishes and drinks from a menu. These are quite typical tasks in many CALL systems. Simpler grammars of ‘ordered’ phrases defined by parallel sentence syntaxes are also used with blocks F and G.

Finally, an unconstrained word-loop grammar was generated for Block C that allows any sequence of words from the vocabulary to be spoken. The set-set perplexity on this grammar is very high, at 340 and the task is therefore not a target exercise type, however it was included for comparative purposes.

To summarise, the grammars and text blocks combine to give recognition tasks which vary in vocabulary size and test set perplexity. These were designed to fall within the user-specification limits of perplexity 10 and active vocabulary of 100-200 words. Statistics for each task are shown in Table 5. Each test consists of speech data collected from 50 speakers except for block F which was recorded by 45 speakers.

BLOCK	LATTICE	#SPKR	VOCAB	#UTT	UNIQ SENT	TEST PERP	AV LEN	MIN LEN	MAX LEN
C0	WLOOP	50	168	1098	23	340	16.9	7	35
D1	MINPAIR-30	50	62	1698	30	3.972	5	5	5
D2	SYNTAX	50	32	850	15	7.457	5.5	2	11
D3	SYNTAX	50	26	194	6	4.080	10.8	9	13
D0	PSENT	50	164	4549	93	2.584	4.8	2	13
E0	PSENT	45	286	2828	63	1.751	7.4	3	13
F0	SYNTAX	50	58	451	10	4.163	15.1	11	21
F0	ORDER	50	58	451	10	2.158	15.1	11	21
F0	PSENT	50	58	451	10	1.165	15.1	11	21
G0	SYNTAX	50	54	550	11	3.008	10.8	8	14
G0	SYNTAX-20	50	63	550	11	3.179	10.8	8	14
G0	SYNTAX-40	50	83	550	11	3.389	10.8	8	14
G0	SYNTAX-60	50	103	550	11	3.519	10.8	8	14
G0	ORDER	50	54	550	11	2.158	10.8	8	14
G0	PSENT	50	54	550	11	1.248	10.8	8	14

Table 5. Recognition Test Sets

3.2. Baseline Recogniser

All experiments that follow use continuous-density Hidden Markov Models (HMMs) trained for 41 speech phones and 2 phones representing silence. Entropic's HAPI 2.0 medium vocabulary decoder was used with a British English pronunciation dictionary.

The baseline system uses speaker-independent acoustic models trained on Southern-British speakers from the WSJCAM0 corpus [8]. The first set of experiments investigate what level of performance can be expected with the non-native speakers using British English context-independent models on the low-perplexity tasks.

Further experiments investigate how accuracy scales as the perplexity increases across the tasks, and what effect the learner's mother tongue and proficiency has on recognition accuracy from native models.

In state-of-the art recognition of fluent, native speech, contextual effects such as assimilation and coarticulation are modelled by using 'triphones'. Triphones model a central phone in the context of its neighbouring right and left phones. Cross-word models allow the neighbour to come from an adjacent word whereas word-internal models may only have triphone contexts within the same word. Context modelling can improve recognition performance considerably if the speaker is native and fluent, however triphones may be less optimal for disfluent native speakers. The effect of context is investigated and an optimal model set is found for recognising the intermediate level speakers.

4 Low Perplexity Recognition with a British Recogniser

The following experiments investigate to what extent a standard British English recogniser can be used within the ISLE system, and how it is effected by speaker proficiency, non-native accent, grammar perplexity and vocabulary size. An optimal recognition system is found and final results produced for each of the target tasks.

4.1. Baseline Recognition and Perplexity

The baseline recognition results were obtained using context-independent monophone models with 16 Gaussian mixture components. A wide search beam of 500 was used to reduce errors. Table 6 shows results for each task ordered with increasing perplexity, from the parallel sentence task to the D2 syntax task.

Task	Perplexity	Avg. Word Accuracy
F P	1.165	100
G P	1.248	100
E P	1.751	99.94
G O	2.158	96.71
D P	2.584	92.97
F O	2.158	94.4
G S	3.008	93.97
D1 M	3.972	90.02
F S	4.163	91.9
D3 S	4.08	86.78
D2 S	7.457	59.26

Table 6. Baseline monophone recognition versus perplexity

Word accuracy falls linearly as perplexity increases, from the parallel sentence tasks to the order tasks to the syntax tasks. The block D results are the exceptions to this showing that phonetic confusions add another dimension of complexity to the task that isn't measured by test set perplexity. The parallel sentence tasks yielded 100% on the tasks where phonetic confusion was limited. Tasks for Block F Syntax and Block D3 Syntax have similar perplexity, 4.08 and 4.16 respectively but the word accuracy differs from 86% to 92%. This is mainly due to the different length constraints, as D3 Syntax allows unlimited length and therefore allows the recogniser more freedom to insert and delete words. The D2 task is the hardest task in terms of perplexity and the result suggests that this kind of task is beyond the scope of the ISLE system with the current baseline system.

4.2. Mother Tongue

Italian speakers perform worse than the German speakers. The difference is almost 6% absolute word accuracy when using monophones on the syntax tasks. The German results on the other hand are comparable to the native speakers on these tasks. The word-internal triphones show a larger distinction between these two groups than the monophones. However it is difficult to draw too many conclusions about the native speakers as the test set is so small.

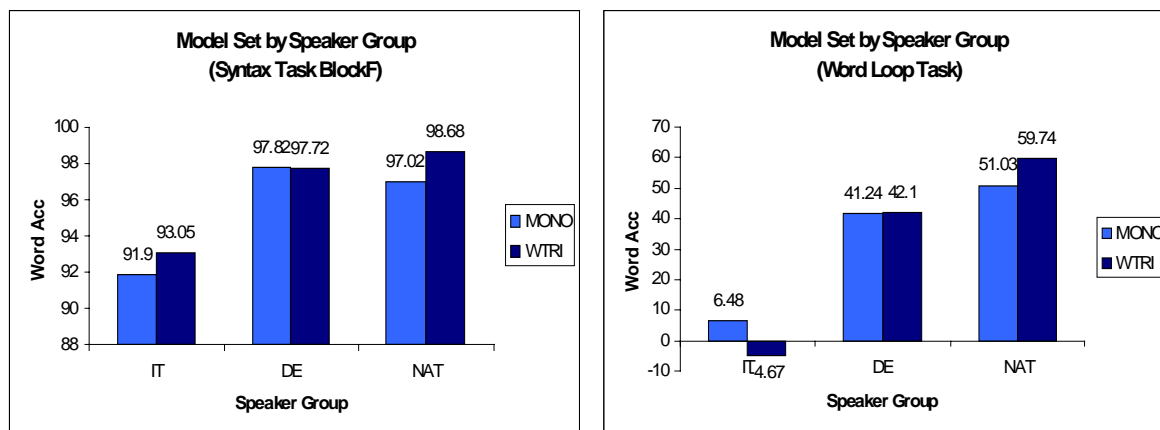


Figure 1. Baseline recognition results comparing speaker groups and effect of context

In order to see more disparity in the speaker groups we can look at the results from the unconstrained word-loop task. In this very unconstrained task, word insertions and deletions occur freely, as seen by the negative word accuracy gained with Italians and word-internal models. This can be explained in part by the common addition of a word-final schwa by Italian speakers, which may be mis-recognised as the short English word ‘A’. The German speakers still perform much better than Italians, but not as well as the native speakers on this task. These results may indicate that high perplexity tasks will be required for more proficient speakers, or for non-native languages that are closely related to the target language such as German.

4.3. Language Proficiency

The plots in Figure 3 map show how recognition accuracy is distributed across proficiency levels (see section 2.2) for German and Italian speakers.

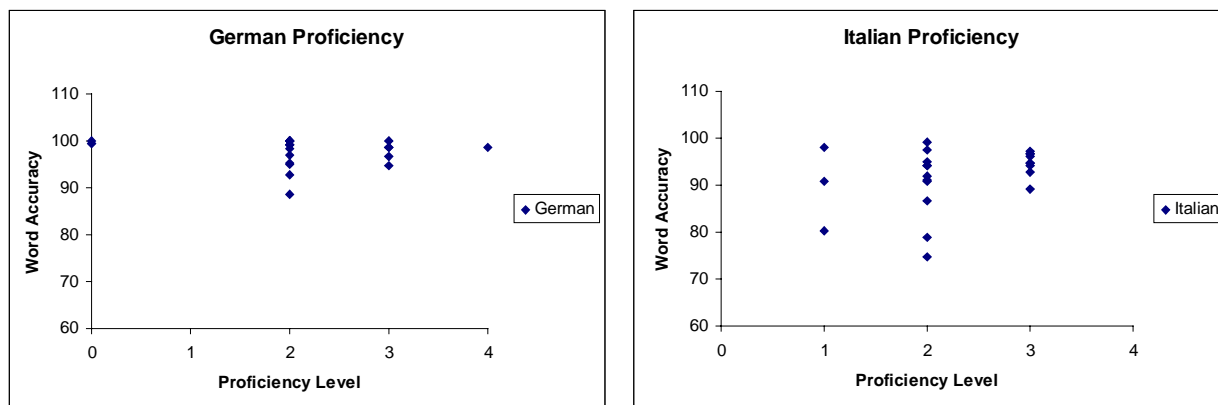


Figure 2. Distribution of Speaker Proficiency (Block G Syntax)

Level 0 represents unknown proficiency, level 1 for beginners, 2 for intermediates, 3 for advanced and 4 for native speakers. There is a larger distribution of recognition scores with the lower proficiency ratings, as some less proficient speakers obtained high accuracy. As the rating increases there is a tendency for the distributions to narrow and centre closer to the top end of the accuracy scale. Comparing the two groups of intermediates, the bottom of the German range is 88% word accuracy, and the bottom of the Italian range is 75%. This shows that the recognition scores agree with the expectations highlighted in section 2.2 that Italians find it harder to adjust to the rules of English phonology and syntax than Germans. There is also a

possibility that the Germans under-rated their proficiency. This will be clarified when the professional accent judgements are complete.

4.4. Acoustic models

Context-independent models have often been used for recognition of elementary non-native speech [17]. This may be because the contextual effects observed in fluent native speech are not mirrored in disfluent learner speech. The monophones are broader models and do not make the assumption that the contextual variation of native speech is equivalent to the variation found in learner speech. The aim of this experiment was to investigate whether this assumption is true, or if context-dependent models can be used in the ISLE system. The optimal number of model parameters is investigated as increasing the number of parameters may improve the modelling of spectral distributions found in native speech but reduce the match with non-native speech. This is investigated by comparing monophones word-internal models and cross-word models.

Systems of increasing complexity were produced by increasing the number of Gaussian mixture components within each model state. Table 7 shows how the number of model parameters increases with increasing mixtures in the context-dependent triphone models. The number of physical models is shown in Table 7, where the ratio of parameters to physical models is given by the *Pratio* and *Params* columns.

Context	WINT		XWRD	
#Phys	3410		4068	
#Mixtures	Pratio	#Params	Pratio	#Params
2	1.073	3660	1.077	4382
4	2.147	7320	2.154	8764
6	3.220	10980	3.232	13146
8	4.293	14640	4.308	17528
10	5.367	18300	5.386	21910
12	6.440	21960	6.463	26292

Table 7. Comparison of Model Parameters

In absolute terms the cross-word models contain more models and more parameters than the word-internal models with the same number of mixture components. This is because there are more contexts to model, however the number of parameters relative to models was designed to be equal. The triphone systems can be compared by looking at models with equivalent *Pratios* (W12 and X12), or equal number of parameters in absolute terms (X10 and W12).

In both cases the word-internal models perform better than the cross word models. Figure 3 shows average word accuracy for Italian and German groups on the block G syntax task and block D minimal pair task.

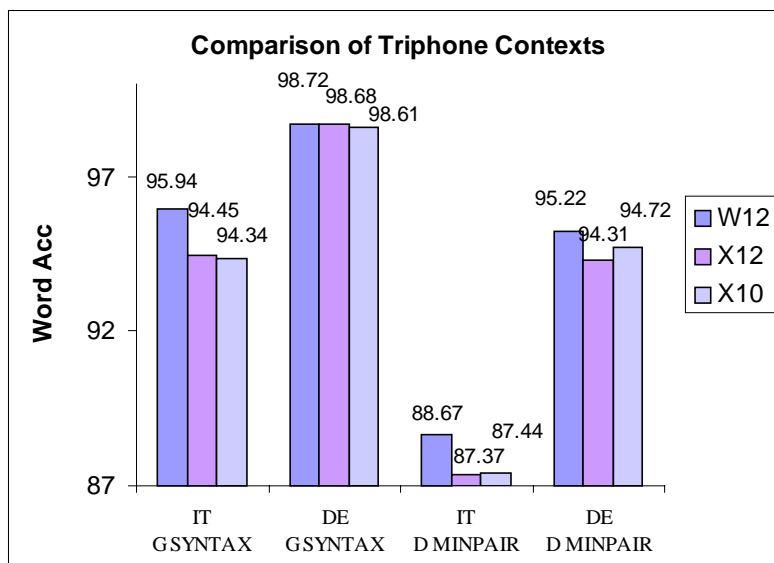


Figure 3. The Effect of Cross-Word Context

Non-native speakers who are disfluent tend to insert more pauses between words and therefore don't display the same cross-word boundary phonological effects. German speakers tend to insert glottal stops between vowels in cross-word contexts whereas English speakers tend to create diphthongs or make use of linking [r], for example in the sentence "It's Diana again" where there is no orthographic 'r' between Diana and again. The Italian word-final schwa insertion that was mentioned in section 4.2 may also reduce the fit of cross-word models.

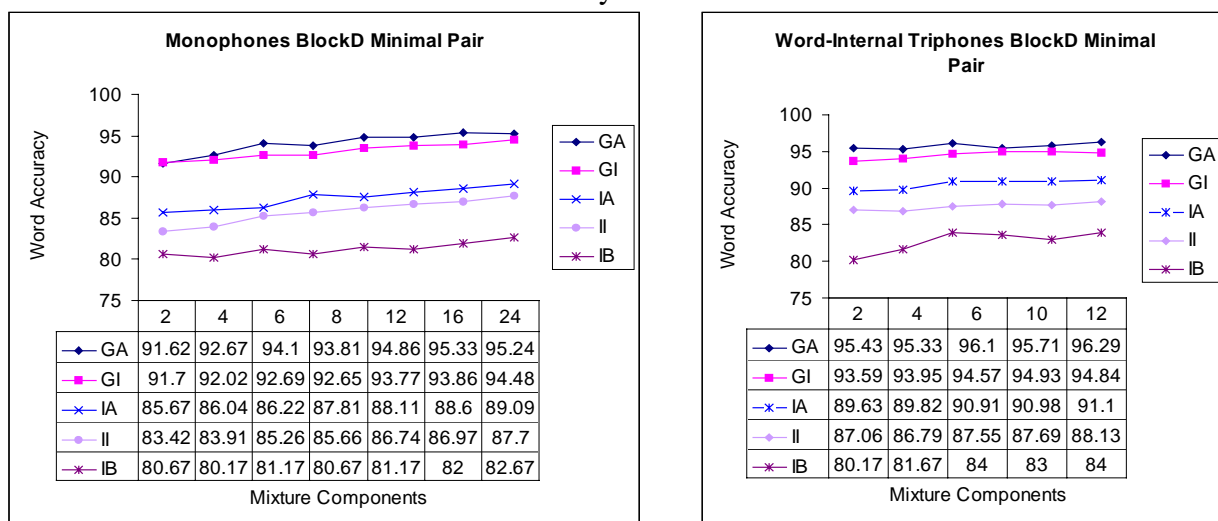


Figure 4. Acoustic Models versus Language Proficiency

Next, the word-internal models were compared against the monophones. The word-internal triphones performed better on all tasks. This suggests that non-native coarticulation within words is quite similar to that of non-native speakers. Figure 4 shows that this is consistent for speakers of all levels.

The results also show that increasing the number of Gaussian mixtures tends to improve accuracy. Performance continues to improve up to 24 mixtures for monophones. Results for the triphones in Figure 4 show that performance starts levelling off at different points with different speaker groups. However this point is largely task dependent and was not shown to correlate with perplexity.

In order to find the best model sets to use in the ISLE system, the scores for each model set were averaged over three tasks (Block D minimal pair, Block G syntax and Block E parallel sentence) on the intermediate German and Italian data.

The optimal model sets on these target tasks were found to be 10 mixture word-internal models for Germans and 12 mixture word-internal models for Italians.

Context	Italian	German	Native
Monophone	24	24	16
Word-Internal	12	10	12

Table 8. Optimal Acoustic Models for Target Intermediate Learners

4.5. Vocabulary Size

The user requirements specification quoted that the maximum active vocabulary size in any one task would be of the order of 100 to 200. From the task vocabulary sizes shown in Table 5, the tasks discussed so far have used an active vocabulary size of between 54 and 286 words. The larger active vocabulary sizes have only been tested within the simplest tasks, as in Block E parallel sentence. In order to test the effect of vocabulary size on a higher perplexity task the Block G syntax grammar was altered systematically. The original grammar contained 11 holiday destinations with an overall vocabulary of 54 words. The number of destinations was then increased to 20, 40 and 60 destinations. These were tested with the 12 mixture word-internal triphones.

#Destinations	Vocabulary	Perplexity	All Non-native
11	54	3.01	97.34
20	63	3.18	97.01
40	83	3.39	95.98
60	103	3.52	95.97

Table 9. Effect of vocabulary size on word accuracy

Increasing the choices to 40 words causes a significant reduction in accuracy over all the speakers, but adding a further twenty words then had little effect. On this test the vocabulary size should be limited to 60 words to reach 97% accuracy.

4.6. Length constraints

The higher perplexity task grammars contain recursive phrases which allow sentences of unlimited length to be recognised. This can lead to reduction in word accuracy if the speech is a poor match. For example if schwa insertion at word boundaries is common in Italian, the schwa could easily be recognised as the word 'A' in the block d3 syntax task. To counter this, word-insertion penalties can be used during recognition.

4.7. Summary of optimal non-adapted results

The experiments described so far have shown that the recognition scores for different speaker groups agree with the general expectations about the effects of mother tongue on a learner's

pronunciation. The scores correlate with the learner's own proficiency judgements within each language group.

Monophones were shown not to be the best choice of models for recognising intermediate learner speech. The cross-word triphones were also shown to be poor models of the word-boundary contextual effects found in the test speech. Word-internal triphone models gave the best results, with 12 mixtures giving optimal accuracy across various tasks. Final baseline results are shown using these models in Table 4.

LANGUAGE		Italian			German		All Avg.
Specified Task	Task	B	I	A	I	A	
ForcedChoice	E0 PS	100.00	99.6	100.00	99.88	100.00	99.77
MinPair	D1 MP	84.00	88.13	91.10	94.84	96.29	91.47
ForcedChoice	D0 PS	88.13	92.20	96.09	97.56	98.93	95.31
Holiday Phrase	G0 O	94.33	95.31	96.12	98.11	98.58	96.58
Holiday syntax	G0 SX	92.78	92.87	95.72	97.74	98.58	95.61
Matching	D3 SX	84.83	86.63	88.42	94.51	96.01	90.57
Description	D2 SX	37.85	53.83	70.13	82.19	85.71	68.72

Table 10. Final results on target tasks using native acoustic models

In developing the demonstrator system it may be useful to set a minimum word accuracy value that should be observed in any exercise selected. If this was set at 96% word accuracy, then the following exercises could be selected when using this baseline unadapted system: Italians and German intermediate learners could use forced-choice exercises when the alternatives are phonetically distinct. German advanced learners could use forced-choice tasks with a high degree of phonetic similarity as in the minimal pair task, however Italian learners could not. German intermediate learners could use the phrase-order grammars as in the Holiday description task and also the higher perplexity syntax grammar, however Italians intermediates would not be given such exercises as they do not gain 96% on either of these tasks.

It is clear that some improvements in acoustic modelling will be required before Italians can perform any exercise other than the simple forced-choice. Improved modelling of the German accent would also be required before the more complex matching and description tasks can be used.

Adaptation Experiments

The next sections describe a set of adaptation experiments aimed at improving performance on these tasks.

The most obvious way to improve acoustic modelling of non-native accents would be to train new statistical models using speech data from non-native learners in question. This full training method has been used to model accented speech successfully [6] however it was not a feasible option within the ISLE project due to time and financial constraints of collecting training data for every source and target language pair. Rather than full retraining, adaptation techniques can be used to transform the model parameters closer to the non-native accent using much less data. The adaptation data can be collected during an initial enrolment session, and during the exercises themselves. MLLR is a technique that can give good improvements with even very

little data [10]. This is due to the fact that all model parameters can be transformed even though many have not been observed. The unseen parameters are grouped with acoustically similar seen parameters into classes and can in effect share the adaptation data. This technique and is attractive for use within the ISLE system as it would allow a shorter enrolment session more suited to non-native learners. MLLR is also a well-established technique that has been used widely for adapting to non-standard accents and different microphones and environments. Alternatives to adaptation exist where knowledge of the transfer effects from the mother tongue is used to map the model. One promising technique known as Model Merging has recently been developed in [18]. The native target models and source model sets are merged according to a single mapping between the two model sets. This technique assumes that models are available in the mother-tongue, but this is less of a problem for mainstream languages as native speech corpora are generally more widely available than non-native corpora.

The next two sections evaluate MLLR and Model Merging as adaptation techniques and compare results to the baseline recognition results. Conclusions are then drawn as to which tasks are feasible in the ISLE demonstrator.

4.8. MLLR

MLLR uses linear transformations for the current model parameters based on the difference between the models and the parameterised target speech. The source models are aligned to an incoming adaptation utterance and statistics from aligned frames are accumulated for each model parameter.

Data sufficiency is a central issue in building robust speaker adaptive transforms. If too little data is available, a poor transform may be generated. To prevent this, frame observation thresholds are set, above which a transform can be generated. Also sparse data is used efficiently by MLLR as groups of similar models may share data. Each group, known as a 'regression class' forms a node in a hierarchical binary 'regression-class tree'. The top node in the tree contains all the model parameters and splits into smaller and smaller nodes until the 'base classes' are met. If a transform is built for the top node it is known as a 'global' transform. As more data becomes available the tree can be traversed and more transforms can be generated for more specific classes.

MLLR may be implemented in various modes depending on the source of the model alignment and how often the transforms are updated. If the transcription comes from recognition results the mode is referred to as 'unsupervised'. In this case the exact word sequence is unknown and the accuracy of the model alignment depends on the recognition accuracy. On the other hand if the speaker is engaged in a reading exercise such as during enrolment session, the target word sequence is known in advance and adaptation is therefore referred to as 'supervised'. In this case the model alignment is produced from a forced-alignment against the word transcription. The second dimension is 'on-line' versus 'off-line'. During off-line adaptation, statistics are accumulated for an entire speaker before the transform set is built. During on-line adaptation, transforms are updated at intervals using statistics accumulated across a set number of utterances. Each updated transform is used to produce the alignments for subsequent utterances, so that in this way the alignments and transforms should become progressively more accurate. This mode may be useful if the length of a dialogue is unknown, or if some accuracy improvements are required very quickly.

As recognition accuracy has been seen to be lower with non-native speakers, unsupervised adaptation may be unsuitable as the inaccurate alignments may lead to poor speaker transforms. It is therefore envisaged that speakers will participate in a short enrolment session during initial use of the ISLE demonstrator system which will allow transforms to be generated in an off-line, supervised mode. This should ensure fairly accurate frame to state alignments and therefore more reliable transforms.

The following set of MLLR experiments therefore use MLLR in an off-line supervised mode. The adaptation data set consists of 60 utterances of blocks A and B of the Everest text. The experiments investigate how many utterances are required for the enrolment session, and how the amount of data effects the types of speaker transforms that can be built. In particular which kind of transformation matrix should be used, whether a global transformation should be used for all phones, or whether class-based transforms are better, and also how the regression classes are defined.

4.8.1. Transform Type

The model parameters are represented by a set of feature vectors. It is often assumed that these features may be grouped into independent diagonal blocks whose characteristics are not correlated, and that transforms can be generated for these blocks rather than for the full matrix of features without losing accuracy [9]. An experiment was carried out to compare the use of a full transformation matrix versus a three-block diagonal matrix, however it was found that a full matrix gave consistently better performance across tasks, with both monophones and triphones. Figure 5 shows 16-mixture monophone results from the block F syntax task and the block D minimal-pair task. This result supports similar findings by Witt with non-native speech [18]. This may be due to the large mis-match in spectral characteristics of the native models and non-native speech.

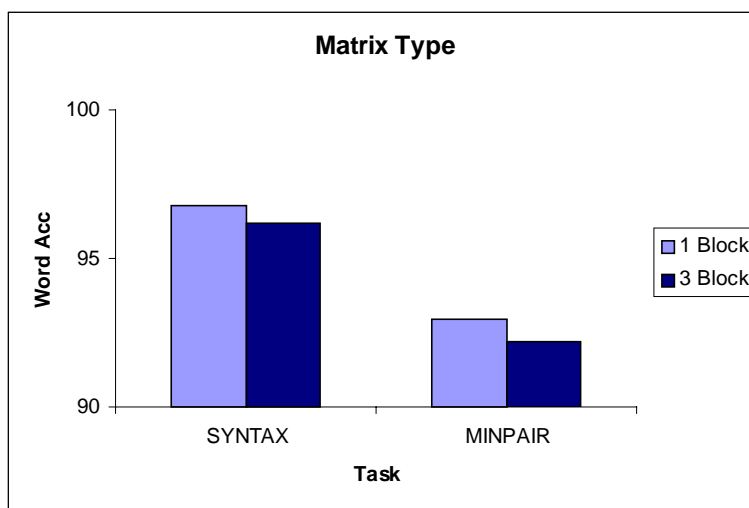


Figure 5. Transformation Matrix Topology

4.8.2. Number of Iterations

If the speech is very poorly matched to the models or the adaptation is unsupervised, the frame to state alignment may be quite inaccurate. Several studies have shown that multiple iterations of MLLR can be used where transforms are generated from gradually better alignments. Witt

[18] found that five iterations were optimal when using a global transform with a full matrix and a word-pair grammar with very elementary non-native students and only five utterances.

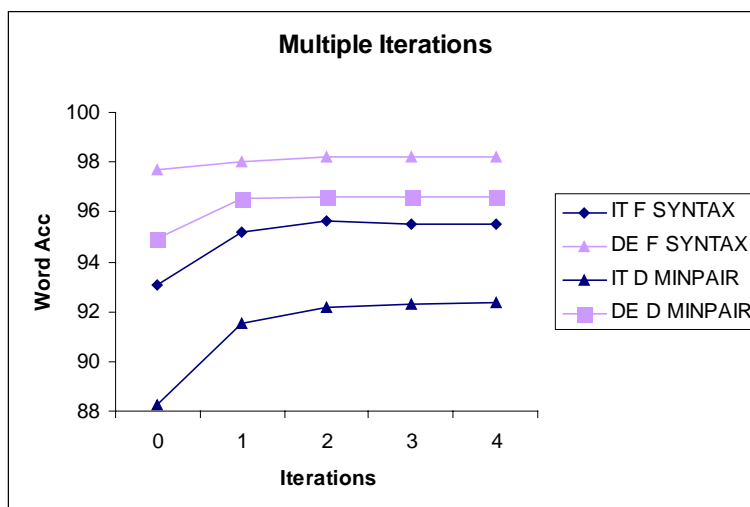


Figure 6. Multiple iterations of MLLR

In Figure 6, results are shown up to four iterations using full-block matrices and 6-mixture word-internal models on two tasks. Iteration 0 denotes the word-internal baseline, iteration 1 shows results from a global transformation and subsequent iterations show results using two transforms, transforming speech and silence model parameters independently.

The first iteration gave the largest gain in accuracy, and results start to level off after just the second iteration. The difference between the results in [18] and these may be explained as the tasks here are more constrained and the students are more advanced speakers plus more adaptation data is available.

4.8.3. Data Sufficiency

As mentioned above, sufficient data must be available for robust estimation of transforms. This is controlled by using frame observation thresholds for regression classes. As more utterances are available, the average number of frames of speech increases as shown in Table 11. If we require 1000 frames per regression class then we would be able to generate 8 transforms given 10 utterances of the “Everest” text and 16 classes given 20 utterances. This is assuming an even distribution of data amongst regression classes which is almost certainly not the case, however it gives an approximate idea of the number of classes to use.

Utterance	Total Frames	Av. Frames per speaker	Classes of 1000
10	401271	8025	8
20	809810	16196	16
30	1151970	23039	23
40	1488526	29770	29
50	1887717	37754	37
60	2302966	46059	46

Table 11. Total frames with increasing utterances

To test the effect of increasing data and the number of transforms, regression class trees were built with a set maximum of 4, 8 or 16 classes. Figure 7 shows recognition results using each of these trees.

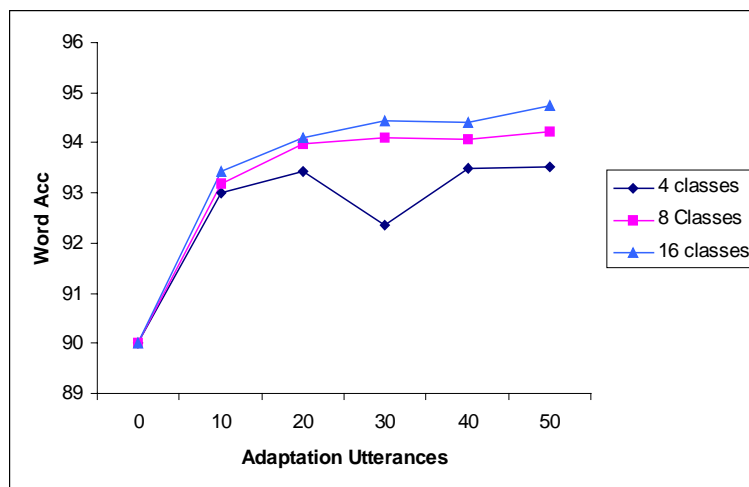


Figure 7. Data sufficiency and regression classes

The tree with 16 classes gave slightly better performance than 4 or 8 classes with 10 utterances, and significantly better performance with 50 utterances as the 4 and 8 class trees reached their transform limits.

4.8.4. Finding Optimal Models for Adaptation

The techniques described in the previous section were combined for this experiment. A full transformation matrix was used with 16 regression classes and an occupancy threshold of 1000 frames. Two iterations of MLLR were carried out as before where the first iteration builds a global transform and the second builds class-based transforms.

Twenty adaptation utterances were chosen as a suitable length for the ISLE enrolment session as recording 60 utterances would be very time consuming. Section 4.8.3 above showed that improvement slowed down after twenty utterances.

An experiment was carried out to decide whether the word-internal models described in section 4.7 should also be used for adaptation.

Model	Base	60 utt	20 utt
Wtri12	95.61	97.49	87.55
Wtri6	95.31	97.02	96.37
Mono16	94.49	97.69	97.22

Table 12. Comparison of Acoustic models for adaptation on

The increased number of model parameters in the 12 mixture triphones meant that 16 full transformation matrices could not be robustly estimated from 20 utterances. This degraded the average word accuracy from 95.61% to 87.55% on the BlockF syntax task. The 6 mixture word-internal models with fewer model parameters gave an improvement from 95.31 % to

96.37% with the same set up. The larger word-internal models did better than the 6 mixture models when sufficient data was available. However, it was found that monophone models produced better non-native adaptation results irrespective of the number of utterances. The 16 mixture monophones gave 97.22 % on this task when adapted with 16 transforms built from 20 utterances. This is a relative reduction in error rate of 49.5% over the monophone recognition baseline and 36.7% over the optimal word-internal baseline.

4.8.5. Accent-Dependent Transforms

So far transforms have been built for individual speakers. Another option would be to build more general transforms for the Italian accent and the German accent using data from the group of speakers. Such a transform set would capture the more general mappings between a single accent and the native source language rather than individual speaker characteristics, and it should be expected that the speaker transforms will be more accurate. On the other hand more data would be available per class.

An accent-dependent transform was built for the 23 Italian speakers using the 16 mixture monophone models with 16 regression classes and all 60 utterances. Results were compared to the speaker-transform results averaged over all Italian speakers.

Task	Italian Baseline	Accent Transforms	Speaker Transforms
D MP	90.02	93.10	94.87
F SX	91.90	96.54	97.28

Table 13. Comparison of accent-dependent and speaker-dependent transforms

The speaker-dependent models do perform better than the accent-dependent models but the accent-dependent models still perform significantly better than the baseline. It should be noted that the test speakers were used within the transform training set and therefore the test is not a true accent-dependent test. However the results do indicate that certain stylistic speaking characteristics or pronunciation errors are shared across the speaker group and can be captured using a general transform set.

4.8.6. Phonetic Regression Classes

The regression classes used so far were defined using an acoustic distance metric to cluster the model parameters. This method assumes that model parameters that are close in acoustic space of native speakers will also be similar in the acoustic space of non-native speakers. This may not be a valid assumption. This experiment investigated using a phonetically motivated regression class tree.

The tree was built with the same topology as the 4-class distance-based tree. The binary tree in Figure 8 starts with a root node containing all model components. This then splits into silence and speech classes. The speech class first divides parameters from consonantal and non-consonantal phones, and the consonantal class then splits into parameters of obstruents and non-obstruent phones.

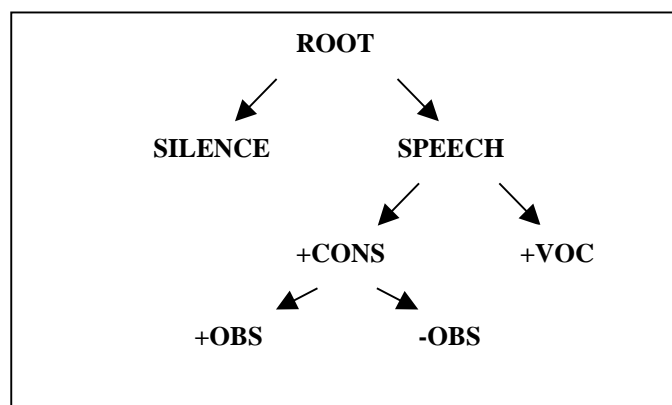


Figure 8. Schematic diagram of phonetically clustered regression-class tree

In this classification the voicing pairs [p] [b][g] and [k] share a transform. Also pairs such as [w] [v] [s] [z] which have the same place-of articulation share a transform. This classification fits many well known pronunciation problems of both German and Italian learners. For example German final devoicing of stops [d][t], German substitution of [v] and [w], Italian fricative voicing [s][z] and Italian vowel centralisation problems such as [ih][iy]. These pairs may be quite distinct in the native acoustic space and might therefore be transformed in separate directions when using a standard distance-based regression tree.

BLOCK	NETWORK	MODEL	BASE LINE	DIST	PHON
D	MINPAIR	WTRI	91.12	93.97	93.97
C	WLOOP	WTRI	17.85	35.19	36.34
D	MINPAIR	MONO	90.02	93.42	93.86
C	WLOOP	MONO	22.95	36.46	38.11

Table 14. Comparison of phonetically motivated and distance-based regression class trees

Table 14 compares the distance-based tree (DIST) and the phonetic tree (PHON) with 16 mixture monophones and 6 mixture word-internal triphones. Results are shown after the second iteration of MLLR, using a full transformation matrix. When used with monophones the phonetic tree performed better than the distance-based tree on all tasks, but when used with the triphones, the phonetic tree and the distance tree performed equally well on the minimal pair task.

The similar triphone results may be due to the way in which triphone states are tied during model training. This is because the tying procedure used decision rules based on the phonetics and phonology of English, the effect of the phonetic regression classes may be distorted. The monophone result on the other hand shows that the acoustic distance may not be the best way to specify which parameters should transform together, and gives an indication that prior knowledge of learner errors may be helpful in guiding class-based adaptation.

4.8.7. Summary of MLLR Adaptation Results

The previous sections aimed to find an adaptation scheme using MLLR for use within the ISLE demonstrator system. The results suggested that an off-line enrolment can be carried out by each new user, during which they read and record about 20 sentences of the “Everest” text. Two iterations of MLLR were found to be sufficient to give a significant improvement, using 20

adaptation utterances to estimate 16 full-matrix transforms. Although the phonetically designed regression-class-tree looked promising, larger regression-class-trees were easier to generate using distance-based clustering so these were used for the final results. Finally, the context-independent monophone models proved to be the most optimal for adaptation even though their baseline recognition was worse than the triphones. Final MLLR results were produced using this set up for all tasks minus the Block E syntax task which was identified as not requiring adaptation.

Speakers	Task	All Speakers			Italian Intermediates		German Intermediates	
		Wtri Base	Mono Base	Mono Adapt	Wtri Base	Mono Adapt	Wtri Base	Mono Adapt
Specified Task								
Minimal Pair	D1 MP	91.47	90.02	95.28	88.13	93.80	96.29	97.44
Forced Choice	D0 PS	95.31	92.97	97.29	92.20	96.25	97.56	98.66
Phrase Order	F0 O	96.58	95.99	97.78	95.31	97.08	98.11	99.11
Syntax	F0 SX	95.61	95.51	97.19	92.87	96.40	97.74	98.90
Matching	D3 SX	90.57	86.78	92.39	86.63	89.56	94.51	94.51
Description	D2 SX	68.72	59.26	84.16	53.83	79.67	82.19	87.17

Table 15. Final MLLR results shown for Italian and German speakers.

Table 15 first shows baseline and adaptation results averaged over all non-native speakers, and then separately for the intermediate Italian and German learners. The baseline results are taken from the 12 mixture word-internal model set and the adaptation results are taken from the 16 mixture monophones results. The averaged baseline result for these monophone models is also given for comparison.

MLLR improves recognition on all the tasks, bringing the average word accuracy for the Block D forced-choice task and both holiday description tasks to above 96%. The accuracy on the minimal pair, matching and description tasks from block D has improved considerably, giving an average increase from 68.72% to 84.16% on the description task, however the word error rate of more than 15% would still result in unreliable diagnosis.

Looking at the individual accents, adaptation brought the minimal-pair task accuracy to above 96% for German learners and also the Holiday ‘syntax’ description task to above 96% for Italian learners, which was the threshold specified in section 4. The speaker adaptation scheme outlined here has therefore widened the scope of viable tasks as hoped. Accuracy increased on already viable tasks such as the forced-choice tasks, which will improve the chances for reliable error diagnosis.

In summary, given speaker adaptation we can expect at least 96% word accuracy on all the ‘essential’ ISLE tasks outlined in the user requirements specification with one exception: the Italian learners performing the minimal pair task. The next section describes a combined approach using model merging and MLLR with the intention of raising the accuracy on this task for Italians.

4.9. Model Merging

The Model- Merging technique [18] is based on knowledge of the differences between the phonological systems of the source and target language, and the idea that learners may try to use

similar sounds from their own phonological system in place of the target phonemes, creating some kind of inter-language.

The model merging technique implements this transfer by finding a mapping the source and target phone sets. For each target phone, the most commonly substituted source phone is found. The mapping is used to merge the two model sets to form a new inter-language model set.

Substitution statistics are found by recognising a set of utterances twice: once using the native source models and once using native target models. The target models are used in a forced-alignment mode to obtain a canonical phoneme sequence according to a pronunciation dictionary. The source models used in a simple phone-loop recogniser. Substitution statistics are gathered by accumulating the number of frames of overlap between each source and target phone pair.

In the experiments used here, the target HMMs were 16 mixture British English monophones, and the source HMMs were 16 mixture Italian monophones. The mappings found are very general and independent of context. Some of these are shown in table 16. These correlate fairly well with the expected phone errors. For example dropping /h/, using /f/ instead of the dental fricative /th/ that gives so many learners of English problems.

English Target Phone	Italian Source Phone
Ae	Eh
Ao	Ow
Hh	Sp
Th	F

Table 16. British English and Italian monophones used for model-merging experiment

Once a mapping is found, the states of the mapped source and target models are merged. This is done sequentially by merging each source model state with its corresponding target model state. Witt used a divergence distance measure to find the closest source mixture component for each of the mixtures in the target state, however in this experiment all mixture components from the source model state were simply added to the target state. The mixture weights of the merged model set were then re-estimated from the same adaptation data.

The mixture weights govern to what extent each HMM models the characteristics of the target or source language. For example, the Italian mixture components should play a more dominant role in modelling speech that is heavily Italian-accented. For these reasons Witt performed the weight re-estimation for each speaker individually. For simplicity however, in this experiment the weights of the merged model set were re-estimated using the combined adaptation data from all the Italian speakers.

The merged models reduced the error rate by 45% over the word-internal best baseline recognition, and by 31% over the full global MLLR transform. However, the model merging result was not as good as the optimal class-based MLLR adaptation scheme.

Wtri Base	Mono Base	Glob. MLLR	Class MLLR	MM	MLLR+MM
86.67	86.90	90.53	94.86	92.86	96.10

Table 17. Comparative MLLR and MM results on Italian minimal-pair task

Following this result an experiment was carried out to combine the two adaptation methods. The merged model set was adapted to each Italian speaker individually as described in section 4.8.4 to produce a new set of speaker-dependent transforms. These transforms outperformed the UK-only transforms significantly on the minimal-pair task as shown in Table 17. The combined approach reduced the error rate by 24% relative to the MLLR only result.

	Wtri 12 Baseline	Mono 16 Baseline	Mono Adapt	Mono Merge	Beg.	Int.	Adv.
MinPair	88.67	86.90	94.87	96.10	95.17	95.51	97.26
ForcedChoice D	93.09	89.20	96.92	97.57	96.04	97.70	97.95
Restaurant Phrase	95.52	94.40	97.69	98.30	97.16	98.57	98.34
Restaurant syntax	94.01	92.15	97.28	97.79	96.65	97.76	98.18
Matching	87.06	81.49	90.34	91.16	89.74	91.03	93.10
Description	57.69	44.53	80.83	81.75	81.07	80.08	84.26

Table 18. Italian results using MLLR and model-merging

Table 18 shows the combined adaptation results broken down by proficiency. The advanced Italian speakers achieve above 96 % on the minimal-pair task but the intermediates still fall just short of 96%, even with 60 adaptation utterances. Despite this, the combined technique gives an overall reduction in error rate of 11% across tasks compared to only MLLR.

5 Conclusions and Recommendations

The goals of this paper were to provide guidelines for the design of exercises within the prototype ISLE demonstrator such that high accuracy will be achieved with non-native speakers performing typical low perplexity tasks. A range of exercise types were evaluated in order to define the upper limits on grammar perplexity and vocabulary size, and to understand how these parameters should scale when the system is used with speakers of different proficiency and mother tongue.

The initial recognition experiments evaluated performance of a standard British English recogniser with no speaker adaptation. It was found that the baseline recogniser could achieve sufficiently high accuracy on low perplexity reading and forced-choice task exercises as long as the grammar perplexity is under 2 and the vocabulary does not exhibit high degree of phonetic similarity. In all other cases some form of speaker adaptation will be required. The best acoustic models for use with this baseline recogniser were found to model word-internal contexts, showing that similar rules for coarticulation within words may exist in accented English and native English.

More complex tasks such as the phonetic minimal pair tasks were evaluated using different adaptation schemes. MLLR adaptation using 16 full class-based transforms was shown to reduce the mismatch between the non-native accent and the British models. It was shown that the manner in which the regression classes are chosen effects the adaptation performance. Classes created using distance metrics were less optimal than classes built using phonetic knowledge of the source and target languages. The result showed that distances in the acoustic-phonetic space of the target language do not necessarily correspond to distances in the acoustic-phonetic space of the source language.

A Model Merging technique was also evaluated using data-driven mappings between the phonetic space of the source language and the phonetic space of the target language. The combined approach of MLLR and Model merging gave the greatest improvements in accuracy with Italian accented speech. In this combined scheme regression classes were selected from some kind of 'inter-language' model rather than the UK model parameters. Investigations into data sufficiency suggested that an ISLE enrolment session of twenty utterances would be suitable, although performance could be improved slightly with more utterances.

The improved acoustic modelling produced by this final scheme meant that more diverse exercise types became viable. Final results showed that German intermediate speakers and Italian advanced speakers would be able to carry out all the essential tasks outlined in the user requirements specification but Italian intermediate speakers were still recognised with less than 96% word accuracy on the phonetic minimal-pair task.

In general the recognition scores correlated well with linguistic expectations of different speaker groups, taking into account differences in the phonological systems of the source and target language. Such expectations can be used to guide the design of tailored exercises within the ISLE system.

Length constraints proved to be an important factor in the design of exercise grammars. Tasks that allowed responses of unconstrained length never reached the target of 96% accuracy even after adaptation. These unconstrained tasks therefore remain beyond the scope of the current ISLE recogniser. Despite this, with clever exercise design, the lower perplexity tasks outlined here can be used to create a many interesting dialogues and goal-oriented spoken language exercises that will create a suitable environment for second language acquisition and will support the accuracy requirements of the ISLE diagnosis component.

Appendix 1: References

- [1] S. Auberg, N. Correa, V. Locktionova, R. Molitor & M. Rothenberg. The accent Coach: An English Pronunciation Training System for Japanese Speakers. In Proceedings ESCA StiLL. pp103-106, May 1998.
- [2] C. Chapelle. Multimedia CALL: Lessons to be learned from Research on Instructed SLA. In *Language Learning & Technology* Vol. 2, No. 1, pp 22-34. No. 1, July 1998: <http://polyglot.cal.msu.edu/llt/vol2num1/article3/index.html>
- [3] M. Swan & B. Smith (eds), *Learner English*. CUP 1987.
- [4] R. Delmonte, M. Petrea, C. Bacalu. SLIM Prosodic module for Learning Activities in a Foreign Language. In *Proceedings Eurospeech*, Sept. 1997.
- [5] F. Ehsani. Air traffic control task for Japanese (Tech. Rep. No. 7-96), Menlo Park, CA: Entropic, Inc, 1996.
- [6] F. Ehsani, J. Bernstein, A. Nagami & O. Todic. Subaruashi: Japanese interactive spoken language education. In *Proceedings of Eurospeech.*, pp681-684, Sept. 1997.
- [7] F. Ehsani, E. Knodt. Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm. In *Language Learning & Technology* Vol. 2, No. 1, pp 45-60. No. 1, July 1998: <http://polyglot.cal.msu.edu/llt/vol2num1/article3/index.html>
- [8] J. Fransen, D. Pye, A.J. Robinson, P.C. Woodland, and S.J. Young. WSJCAM0 corpus and recording description. Technical report CUED/F-INFENG/TR192, Cambridge University Engineering Department, Cambridge, U.K., 1994.
- [9] A. J. Leggetter. Improved Acoustic Modelling for HMMs using Linear Transformations. Phd thesis, Cambridge University, 1995.
- [10] A. J. Leggetter and P. C. Woodland. Flexible Speaker Adaptation for Large Vocabulary Speech Recognition. In *Proceedings of Eurospeech*, pp1155-1158, Sept. 1995.
- [11] N. Nagata. Input vs. Output Practice in Educational Software for Second Language Acquisition. In *Language Learning & Technology* Vol. 2, No. 1, pp 23-40. No. 1, July 1998: <http://polyglot.cal.msu.edu/llt/vol2num1/article3/index.html>
- [12] L. Newmeyer, H. Franco, M. Weintraub, & P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. *International Conference on Spoken Language Processing*, Sept., pp1457-1460, 1996.
- [13] L. Newmeyer, Franco, A. Sankar, & V. Digikalis. A comparative study of speaker adaptation techniques. In *Proceedings of Eurospeech*, pp1127-1130, Sept. 1995.
- [14] M. Rypa. VILTS: The voice interactive language training system. In Proceedings of CALICO, July 1996.

- [15] B. Sevenster, G. deKrom & G. Bloothoof, Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs. In Proceedings ESCA StiLL. pp91-94, May 1998
- [16] M. Swan & B. Smith (eds), *Learner English*. CUP 1987
- [17] S. Witt & S. Young Language Learning based on non-native speech recognition. In Proceedings of Eurospeech pp 633-636, Sept. 1996.
- [18] S. Witt Use of Speech recognition in Computer-assisted Language Learning. Phd thesis, Cambridge University, 1999.