



ISLE

## Pronunciation Training: requirements and solutions

Project: LE4-8353  
Deliverable: D1.4

Version	4
Date	29.1.1999

**ISLE Deliverable**

Project Number	LE4-8353
Project Title	Interactive Spoken Language Education [ISLE]
Deliverable Type	RE
Distribution	P
Deliverable ID	D1.4
Expected Delivery Date	T7
Actual Delivery Date	26.1.99
Title of Deliverable	Pronunciation Training: requirements and solutions
Author(s)	Eric Atwell (Leeds University), Dan Herron (Hamburg University), Peter Howarth (Leeds University), Rachel Morton (Entropic), Hartmut Wick (Ernst Klett Verlag)

OT	RE	SP	PR	TO
Other	Report	Specification	Prototype	Tool

C	P	R
Consortium	Public	Restricted

**Revision History**

Version	Date	Status	Author(s)
1	14.12.98	Draft	Leeds University [Eric Atwell] et al
2	15.12.98	Draft	Leeds University [Eric Atwell] et al
3	26.1.99	Near final	Leeds University [Eric Atwell] et al
4	29.1.99	Final	U Hamburg

## **SUMMARY**

This report summarises the main findings of ISLE Workpackage 1, “Requirements Specification”. Specifically, this report includes selections from three internal reports delivered for Workpackage 1, covering ISLE user requirements and proposed solutions:

- The **requirements** of prospective users of ISLE are investigated in Report D1.1, “User requirements specification”; and in Report D1.3 “Market Analysis”. The former report provides ISLE partners with guidance on what users need in a computer system for pronunciation tutoring in EFL. It presents the results of a variety of data-gathering exercises (questionnaire surveys, software evaluation and literature survey), with interpretation and analysis. The main conclusions relate to the model of spoken English the system should use, the kinds of learning activity it should provide and the type of feedback it should aim to give. The “Market Analysis” report examines the market for educational software in general and ISLE-like systems in particular, including an analysis of competitors.

- The **solutions** proposed by the ISLE project are outlined in Report D1.2 “Specification of technical parameters”. This report and annexes specify the ISLE system structure, exercise types, ISLE calls available to the high-level interface, and procedures for validation and verification, including collection and annotation of a corpus of learners’ speech.

This report is for public distribution, and is aimed at potential users of ISLE, in particular English language teaching professionals.

**Contents**

ISLE Deliverable.....	2
Revision History.....	2
SUMMARY .....	3
Contents.....	4
1. Introduction .....	5
Part 1: Requirements .....	6
2. Analysis of ISLE user requirements .....	6
2.1 Accents .....	6
2.2 Data presented to users.....	7
2.3 Activities .....	7
2.4 Feedback .....	9
3. Market Analysis .....	9
4. Validation .....	11
Part 2: Solutions .....	11
5. Proposed solution: outline of ISLE technical specification .....	11
6. The ISLE System Architecture .....	12
6.1. Recognition .....	13
6.2. Localization.....	13
6.3. Diagnosis.....	14
6.4. Stress-detection .....	14
7. Exercise Types .....	15
8 Feedback .....	15
9 Corpus-based Verification and Validation.....	16
10 Conclusions .....	18
Appendix 1: Survey of language-learning products with speech recognition .....	19
Appendix 2: Online References .....	21

## **1. Introduction**

The Interactive Spoken Language Education (ISLE) project, funded by the European Commission, aims at introducing speech recognition technology into future Computer-Assisted Language Learning (CALL) products for adult learners of English. One of the most prominent goals is to provide an appropriate level of specific feedback to the student in order to point out possible directions for the improvement of pronunciation. Existing courseware products on the market that use speech recognition capabilities are often developed without direct input from the end-user - for example, the feedback to the student is often restricted to a global quality measure without specific advice on how to improve pronunciation. ISLE aims to improve on this by localising errors to specific words and phones and providing clear feedback to the student (e.g., that a particular phone substitution has occurred, and what the student can do to correct this.)

The research efforts of the ISLE project are concentrated on development in four main areas: (1) an HMM-based recognition engine that is fast, reliable, and robust even when presented with non-native speech; (2) a localisation mechanism that reliably pinpoints deviations in the student's pronunciation from the native model; (3) a diagnosis mechanism that explains pronunciation errors as the product of phonemic or orthographic -> phonemic errors by the student, allowing the system to focus on particular errors and to provide useful feedback; and (4) a lexical-stress detector to locate misplaced stress, which is clearly a problem for many learners of English. Exercises and feedback are to be presented within a high-level multimedia authoring system.

Models adapted to non-native speech are to be used during the recognition stage in order to increase accuracy. When localising and diagnosing errors, however, non-adapted models are used, to maximise the probability of detecting errors. Additionally, although any phone-level error (substitution, deletion, or insertion) can potentially be detected, the system will concentrate on errors that are expected based on the student's first language or that stem from common orthographically-induced mispronunciations; this further increases accuracy and lowers the probability of false alarms.

Integration of these components creates a natural learning environment in which the student is never responsible for self-diagnosis. Besides providing the student with immediate feedback, long-term performance data (at the exercise, phrase, word, and phone levels) is collected to allow the student's performance to be tracked across time. A medium-sized corpus of non-native, intermediate-level English (from Italian and German speakers) is to be collected. This corpus will be transcribed and annotated for phone- and stress-errors and will be used for performance evaluation.

The project will produce a software toolkit with components for spoken language interaction in controlled exercise environments, speech quality assessment, and feedback presentation; these tools will be easily interfaced to other authoring systems, via the OCX-based mid-level interface, which handles all calls to the recognition and diagnostic engines.

In summary, the strategic goal of ISLE is to exploit available speech recognition technology to improve the performance of traditional language learning software. This Report covers the main finding of ISLE Workpackage 1, “Requirements Specification”. At the outset of the Project, we sought guidance on what users need in a computer system for pronunciation tutoring in English as a foreign language. The conclusions drawn from this analysis of user Requirements led us to our Solution: specifications of the ISLE system structure, exercise types, ISLE calls available to the high-level interface, and user-centred verification and validation.

## ***Part 1: Requirements***

### ***2. Analysis of ISLE user requirements***

The major issues anticipated from the investigation of user needs are:

1. the accent(s) of English to be aimed at and/or presented to the user for imitation
2. the speech, text and visual data to be presented to users as stimuli
3. the activities that users will engage in
4. the feedback that will be given in response to the users’ speech

The major sources of data for the investigation of user needs were:

- A questionnaire administered to English language teachers via the following distribution lists:
  - ISLE project participants
  - Staff of the Language Centre, University of Leeds
  - BALEARP members (British Association of Lecturers in English for Academic Purposes) by post
  - EUROCALL subscribers (the European Association for Computer Assisted Language Learning) e-list
  - TESLCA-L subscribers (Teachers of English as a Second Language-CALL) e-list
  - ICAME (International Computer Archive of Modern and Medieval English) conference participants
  - IATEFL (International Association of Teachers of English as a Foreign Language) Pronunciation Special Interest group
  - CAPITAL (Computer Assisted Pronunciation Investigation Teaching and Learning) interest group of CALICO (Computer Assisted Language Instruction Consortium) e-list
- A questionnaire administered to adult learners of English
- Evaluation by language learners and professionals of existing software
- A brief survey of the literature on pronunciation teaching

Details of the user survey are in report D1.3. The main conclusions were:

#### **2.1 Accents**

a) The user questionnaire and the literature survey both produced a wide diversity of strongly held views, ranging from ‘anything but RP<sup>1</sup>’ to ‘RP is a model appropriate for international

---

<sup>1</sup> RP refers to the “received pronunciation” of British English

intelligibility' to 'doesn't matter'. There was an overall preference (even if sometimes reluctant) for something like standard British English as the model for users to aim at, based on educated varieties of southern British English, and possibly also northern and Scottish English. It is important to consider who the users of the program will be communicating with: chiefly with native speakers or chiefly with other non-native speakers. This may affect criteria such as intelligibility in the design of the system.

b) Accents presented to learners should ideally be varied by sex, age, region, with as much choice as possible offered to the user, and information about the speakers. There will of course be a cost to the project in providing variety, in terms of data collection and preparation.

## 2.2 Data presented to users

Of the categories of stimulus and feedback (text, sound and graphics) there is a clear preference for sound (listen and repeat etc), little interest in textual prompts and reservations about the value of wave forms. Some respondents suggested that learners would have difficulty interpreting such visual information, though others thought that animated representations of articulation could be useful. The commonest additional suggestion was for learners to participate in dialogues. This was not included in the list of possible preferences to choose from, as we had already ruled this as beyond the scope of ISLE on grounds of technical feasibility; however, there is clear user demand, which could perhaps be addressed in future research.

Teachers expressed a marginal preference for instructions and feedback in English over L1 (the student's first language) or both languages. Learners showed no interest in the use of the mother tongue on screen. Comments suggest that a choice might be offered and that the more advanced learner would choose English.

The speech data presented should partly be selected to give German and Italian learners practice in their specific problems. The units of speech presented and practised could include:

- single syllables for ear training, perhaps reinforced with video
- production of problem phonemes, again possibly reinforced visually
- The survey showed a clear desire for IPA (or some version of it) to be available as an option, with some training or guide to its use as a reference tool.
- single words for phonemes and word stress
- phrases for stress patterns and weak forms
- sentences for contrastive intonation
- utterances presented with wave forms, orthographic text aligned with the graphics and with some kind of indicator moving along the wave form as the sound is heard (e.g. by the colour changing with the speech)
- texts read aloud
- dialogues

## 2.3 Activities

Two related features stand out as important in the surveys: control and variety. One of the main lessons from the evaluation of *Auralang* is the importance of giving the user the option of active control over the recording phase (i.e. by having a start and stop button). This would clearly have implications for analysis and feedback. Control could also be provided in terms of choosing between accents to listen to, having a choice between scored and unscored practice, size of pronunciation unit to practice, etc.

There are two views on providing as much variety as technically possible. On the one hand, this may be wasted effort if learners don't make full use of the features offered and they might be confused by a complex structure. On the other hand, variety would allow for the diversity of users' preferred learning styles and may encourage repeated use of the program. This issue is best resolved in the light of market information gathered by the publisher partners (D1.3). The proposals presented here assume that maximum variety is desirable, with the obvious caveat of technical feasibility and cost.

A further feature of the system that emerged from the survey is that pronunciation training depends to a large extent on accurate, guided listening. It would be beneficial, therefore, if material were included for the introduction and training in key problem areas: the articulation of individual sounds, word stress, the contrast between strong and weak syllables, weak forms, phrase and sentence stress, and intonation.

The activities and styles of presentation might include:

- Users hear a speech segment and select (with the mouse) on screen from options to indicate correct listening (e.g. choose from different phonetic symbols, or orthographic words, match words with sounds, etc). This would be appropriate for initial ear-training and familiarisation with phonetic symbols.
- Users listen to utterances and see both the text and a wave form with the stressed syllables indicated. This, again, would be for initial training in the identification of word, phrase and sentence stress. A distinct component on recognising weak forms should be provided.
- Users view and imitate short video sequences or an animated mouth for a limited set of individual problematic target sounds.
- Users select, listen to and repeat phrases and longer utterances. The shorter units would focus on articulation and accuracy of individual sounds and word stress, the longer ones for sentence stress and intonation. These can be presented in sound, in ordinary orthographic text and with a wave form in parallel, perhaps with users given the option of which combination they see.
- Users participate in dialogues. Ideally, this would involve a conversation between learner and system on a specific topic, imposing minimal artificial constraint on the user's language. However, current speech recognition technology would have difficulty coping accurately without a highly-constrained syntax. Instead, a system could simulate a conversation. This could be done in the manner used in *Auralang* (users hear one side of the conversation and are presented on screen with options for responding in speech; the system recognises which choice has been spoken, and this determines the next utterance in the exchange). Alternatively, 2 speakers can be presented on screen (with their photographs) and the user chooses which part to play in the dialogue. A lot of control should be given to the user and enough information provided on screen for them to keep track of the dialogue if they choose to repeat, recap etc.
- In all these activities, the facility should be provided for users to listen again to their own speech for comparison with the target.
- Activities should relate to the specific pronunciation problems which learners and teachers consider to be most important. Features related to stress received great emphasis: word stress, weak forms and intonation. Non-native teachers, and learners, place greater weight on problems related to individual sounds than do native-speaker teachers. Prominent among the further comments was the view that pronunciation of individual sounds should be presented contrastively with L1. For German learners of English the major problems are intonation, followed by individual sounds and then stress-related features. For Italians, individual sounds are the biggest problem followed by word stress.

## 2.4 Feedback

The “benchmark” representative existing system we tried out, Auralang, offered limited pronunciation tuition. Users were given a raw score (0-7) and shown a target waveform to compare against their own speech waveform; but users are given no further guidance as to how they should modify specific aspects of articulation to produce a better match to the model. Our survey indicated that effective feedback involves, among other things:

- Users should be able to hear their own speech in comparison with the target and to be given guidance in analysing the difference. This may be done by means of wave forms, in which case they must be accurately aligned. The system should direct users to the significant differences and as far as possible inform them of the type of mismatch, especially the length and quality of individual phonemes.
- A context-sensitive help system could provide additional detailed guidance, perhaps with examples of correct articulation, stress pattern etc.
- The users’ attempts at imitating the target sounds should be stored during a session and evaluated so that they can be reviewed at any time to compare levels of success.
- A clear wish for some sort of scoring was reported, and several teachers commented on its motivating effect for learners. Some respondents had reservations about the problem of “correctness” in pronunciation scoring, and said that an indication of “comprehensibility” would be preferred.
- As much additional information and guidance as possible should be available, partly to allow the learner’s route through the material to be individualised and varied. This would help to make repeated use of the program more stimulating.
- Dialogue exercises could encourage learners to activate their speech, and build confidence; a low “acceptability” setting for the recogniser would allow unstressful production with positive feedback in the form of encouraging conversational responses.

## 3. Market Analysis

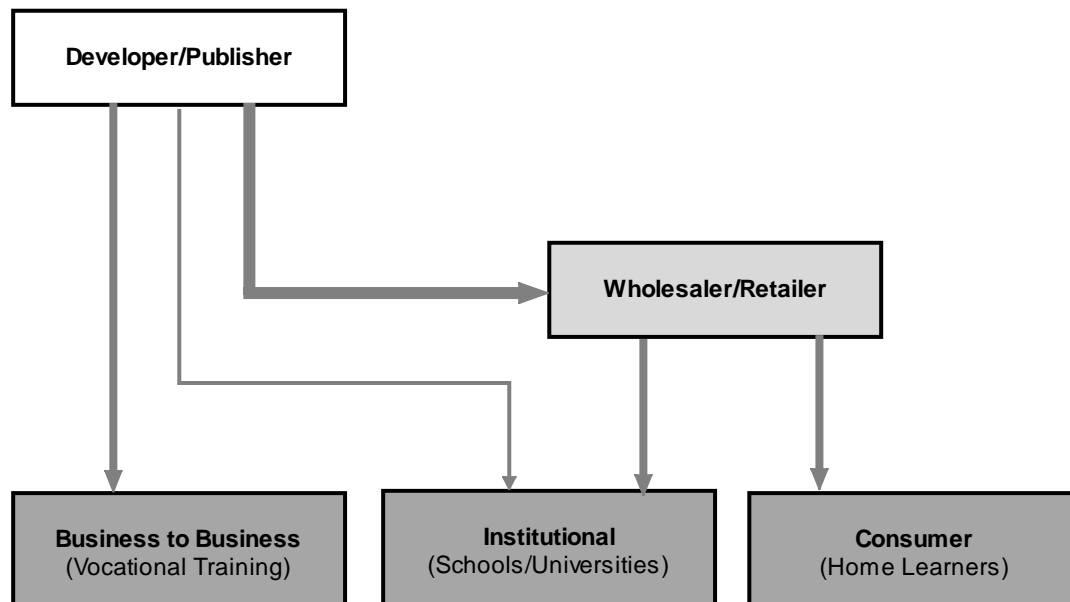
The market for edutainment in a wider sense is well prepared, the largest potential market being in the European Union. Attractive, high-quality educational titles are now needed to penetrate this market. The number of software titles related to language-learning can only be guessed at, and it is thought that these titles represent around a half of all educational software. There has also been an increase in recent years of language-learning software that incorporates speech recognition, but these have often been developed without direct consideration of the special requirements of this domain. They are often fairly basic and offer limited feedback which, according to preliminary experiments, can have a serious negative effect on the student. Deficient language skills are an obstacle to business in most European countries, and both companies and the public authorities are trying to exploit multimedia to combat this problem, it being cheaper to use computers to train staff than to send them to training classes. The market for multimedia software for use at home is also considered to have extremely high potential.

ISLE fills this gap in the market, but must be implemented quickly since competitors who are currently active in the CALL software market and indirect competitors dealing with SR technology could soon become direct competitors.

Generally speaking the market for language learning materials is structured similarly to the market for other educational software. That is, there are the developers that supply the three end-user market segments via trading agents or directly. Depending on the needs and

preferences of each segment the purchasing criteria and consequently the product packages offered vary widely.

#### The Educational Software Market Structure



Today, the market sees an advance of different language learning tools with speech recognition capabilities which unfortunately have often been developed without direct consideration of the special requirements of the application domain. Besides simple recording-and-playback facilities a few programs offer an automatically-derived quality assessment for the student's input (e.g. "Auralang", see user evaluation in Section 2.4). They are, however, rudimentary since they are usually restricted to a fixed set of model utterances (usually a small set of lexical items for lexical testing). Additionally, the feedback offered is of limited use since it only gives a judgement for whole utterances based on a comparison of the input utterance with a recorded one. However, this fails to pinpoint the particular pronunciation problem precisely. Being unable to produce more detailed explanations, the system leaves the student without any indication as to what needs improving. In fact, preliminary experiments revealed a serious negative effect on the student: since he or she is not given a reason for a good or bad feedback, confidence in the system rapidly diminishes and after a few seemingly implausible examples of feedback, the student often just ignores the rating. Generally, an overall didactic framework for teaching foreign language pronunciation is lacking.

The assumptions at the start of the ISLE project are still valid, that is plenty of CALL products are available but they are neither of high quality in providing differentiated pronunciation feedback, nor tailored to specific faults made by learners of specific nationalities. That means that user requirements, as specified above, are still not met by products on the market. Therefore, the ISLE system fills a gap in the market having no direct competitors at the moment. Of course, competitors currently active in the market of CALL software and indirect competitors dealing with SR technologies may eventually become direct competitors. Consequently it is critical to the success to quickly implement the ISLE developments into real-life software packages as demonstrators, and to licence the ISLE SDK to third party publishers in order to quickly reach a "critical mass" for successful market penetration.

In order to achieve this goal ISLE are focusing their marketing strategy on the ISLE SDK. As laid out in the target market analysis this is the most promising option. The sale of the prototypes and runtime licences is a by product carried out by the publishers within the ISLE consortium and additionally by future customers after they have purchased and used the ISLE SDK.

#### **4. Validation**

One remaining “requirements” issue is the need to test and evaluate the system. During development, system components must be tested using off-line test data: examples of possible user inputs, collected in advance. The initial and final prototypes can also be subjected to large-scale off-line testing, using a Corpus of English language learners’ utterances. Unfortunately, there is no existing Corpus resource available (e.g., from an archive such as ICAME [International Computer Archive of Modern English] or ELRA [European Language Resources Association].) For a few specific problems such third-party corpora might prove helpful, but a true validation requires that we collect a small corpus of non-native English annotated to the proper degree.

However, our needs are very specific: for verification and validation, we need a test Corpus of learner’s spoken utterances typical of the demonstrator’s prompts, spoken by typical examples of our target learner groups (German and Italian intermediate and advanced learners of English). Pronunciation errors in the utterances should be marked or annotated, so we can then see how the system fares with known errors. Therefore, we must collect and annotate our own. We will thereafter be able to make this resource reusable by others via an archive like ICAME or ELRA.

We must also expose it to on-line verification: user trials. For on-line validation of the prototypes, we need access to users – teachers and learners – who can try out the system and give us feedback. In disseminating and promoting the final system to English language teachers and learners, a qualitative measure of system match to user requirements is at least as important as corpus-based quantitative measures such as error-detection rates. Furthermore, we do not have enough time, money, users, or exercises to attempt a large-scale investigation of quantitative metrics.

### **Part 2: Solutions**

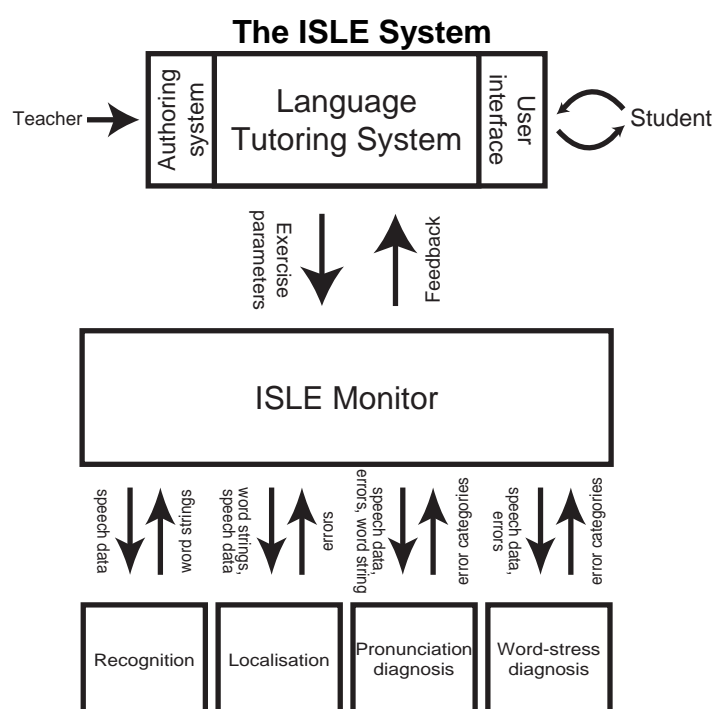
#### **5. Proposed solution: outline of ISLE technical specification**

The strategic goal of ISLE is to exploit available speech recognition technology to improve the performance of traditional language learning software. ISLE technology adds diagnostic components to the underlying speech recognition system in order to determine where, when and which pronunciation mistakes have been made; it also gives the end-user feedback about how to compensate for or correct these mistakes. The core of the ISLE system is based on four different components: a speech recognition module, an error localization module, a diagnosis module for mispronunciations and a diagnosis module for word-stress errors.

The principal ISLE product will be a self-contained system that can be used by many high-level packages in order to add speech-recognition and diagnosis capabilities. In order to be successful, it will come with high-quality documentation: an installation and integration handbook, guidelines for the generation of new applications, and guidelines for the design of new exercises for integrated courseware packages. A prototype system for individualised training will be developed, available in two versions: English pronunciation for German learners and English pronunciation for Italian learners. In the prototype, the student can interact with the system by reading a given text, by matching items, through multiple-choice exercises, and through exercises on the pronunciation of minimal pairs (e.g. *dip* and *deep*). The system evaluates the input and gives feedback about the location and type of mispronunciation, and about how it could be improved. The feedback is presented in several different forms: scoring, playback and textual transcription of an utterance, presentation of model pronunciations, highlighting of mispronunciations in the transcription, giving a textual description of the error.

## 6. The ISLE System Architecture

The ISLE diagnostic components are capable of determining when, where, and which pronunciation mistakes have been made. ISLE also introduces a feedback module that will effectively inform the end-users of what their pronunciation mistakes are and how they can compensate for or correct them. Developers can integrate the full capabilities of the ISLE System directly into their programs using authoring systems such as Asymetrix's *ToolBook* and Macromedia's *Authorware* or development environments like Microsoft's *Visual Basic* and *Visual C++* because the system is based on the Component Object Model (COM) architecture under Win32.



The top-level of the ISLE demonstrator system will be implemented in Asymetrix ToolBook, a high-level multimedia scripting language. A small-scale pronunciation-training system will be developed, which includes:

- 1) tracking of individual students
- 2) various exercise types (e.g., question-and-answer, reading exercises, etc)
- 3) detailed examples of feedback for various errors
- 4) long-term progress reports for each student, showing improvements across time

The mid-level interface between the high-level component (which in theory can be built by any teacher/multi-media designer with very little knowledge of the underlying components) is built on Microsoft's OCX/COM technology. This technology allows the critical and complex low-level functions to be encapsulated so that the top-level interface (implemented in any of the languages or packages that can handle OCX) can deal in a consistent and simple, yet still fast and powerful way.

The ISLE project will release, along with the actual software, detailed examples of how to build an interactive pronunciation training exercise, showing the necessary calls to the mid-level. Essentially, though, this involves a set of initialisation calls, calls to specify properties of the individual student (e.g., his original language), calls to prepare for a new exercise, diagnostic calls, and clean-up calls. Although modest programming ability will still be required, this will greatly simplify the task of anybody creating a new exercise.

The low-level core of the ISLE system is based on four different components:

### 6.1. Recognition

The state-of-the-art IHAPI speech recognizer from Entropic is responsible for recognition. IHAPI provides a detailed and complete set of routines that allow for all aspects of recognition—working with audio inputs, specifying what is to be recognized, querying the recognizer for the results of each recognition, etc. This encapsulates the recognition components, allowing them to be modified and upgraded without affecting other components.

IHAPI uses Hidden Markov Model technology in to implement the recognition; essentially this means that a large corpus of speech data is transcribed and 'models' of each speech-sound (known as the "phones", and roughly equivalent to the phonemes of a language) are created. Thus extensions of ISLE to languages other than English require only that new models be trained (which has, in fact, already been done for many languages.)

### 6.2. Localization

In order to determine which parts of a student's response contained errors or poorly-pronounced words or segments, several routines examine the results of the IHAPI recognition, apply user- or system-controlled thresholds, and return lists of words or sub-word segments that should be 'flagged' as having been poorly pronounced. This is achieved essentially by querying each phone model as to how well it 'fit' the region of speech identified as containing that phone.

A key point of the ISLE system, however, is that students' utterances should always be recognized, even when they are poorly pronounced. This leads to a conflict, however, in that the recognizer should be especially 'generous' in the recognition phase, allowing the student to make many errors; yet it should be rather strict in the localization phase, forcing the student to pronounce the words properly, even to the degree of discriminating between a British or an American accent. To cope with this conflict, the ISLE system will use different sets of models in the recognition and localization tasks. Those for recognition will be trained on a wide-

ranging set of pronunciations and will be ‘adapted’ somewhat to the student’s speech patterns; while those for localization will be trained on a much narrower range of native accents.

### 6.3. Diagnosis

The most innovative and important part of the system requires a determination of *how* the student mispronounced a word or phoneme. The ISLE system will implement this with several strategies. Every error will be considered to stem from:

- 1) an error predicted by the student’s first language (e.g., a German student might mispronounce “wine” as something like “vine”)
- 2) an error predicted by the complex and contradictory rules of English orthography and pronunciation (i.e., a student might mis-read “dough” to rhyme with “cough” or mis-read the verb “live” as the adjective “live”)
- 3) any other inexplicable error

Each of these kinds of diagnoses will require different forms of feedback in order to best help the student. Errors of the second class are fairly easy to detect, as they require simply looking for certain (incorrect) pronunciations of the words in the sentences. Errors of the first type, however, will in general be more difficult to handle, since they may often be the result of a substitution of a phone from the student’s native language for the proper English phone. Ideally this would be dealt with by training models on *both* German and English speech data, so that the ISLE diagnostic components could look for words in which German, rather than English, phones are used. While this is technically possible (although challenging), it requires training data that is not currently available. Fortunately, at least for the cases of German → English and Italian → English, the native English phone set contains almost all of the phones from German or Italian that a learner might mistakenly use.

Successful diagnosis of an error means, in this view, finding a rule that accounts for the pronunciation. It is then fairly easy to provide feedback to the student. In the early ISLE demonstrators this may be something that is correct but pedagogically less useful (e.g., a message such as “you said ‘v’ instead of ‘w’”), but it is simple to map from errors to more useful messages, and then to allow the student to practice that word or that phone.

### 6.4. Stress-detection

English is complex not only in pronunciation, but also in placement of accent; thus the final component will detect errors in word-stress (e.g., the student mis-stresses the verb “CONtrast” as the noun “conTRAST”.) This is accomplished by using the word- and phone-level segmentation provided by the recognizer together with the acoustic properties of the speech wave, in order to make guesses about which syllables were stressed. Although word-level stress is very easily heard by humans, it is quite difficult to do so automatically. The ISLE approach will help to overcome this difficulty by integrating the phone-level information from the recognizer to take advantage of stable correspondences in English between vowels and stress (i.e., it is impossible to have a stressed schwa, or “uh” sound.)

Stress feedback will, of course, require knowledge not just of the student’s stress pattern but also of the *correct* stress pattern. By comparing the two, the mid-level will be able to pass back to the top-level pointers to any words with incorrect stress (and to point out what the correct and incorrect patterns are). The top-level can then provide feedback, which is fairly direct for stress errors (i.e., “the word ‘analyse’ should be stressed on the first syllable; you stressed the last syllable.”)

## 7. Exercise Types

The immediate result of the ISLE project will be a prototype system for an individualised training of pronunciation capabilities for foreign language learning. Our choice of exercise types must be a compromise between user preferences as set out in Part 1, and what is technically feasible in the prototype system as specified above.

The prototype system will accept spoken input in a controlled environment with the following **exercise types**:

1. Reading a given text (possibly presented as part of a simulated dialogue)
2. Matching items (oral combination of suitable items from several lists, such as choosing the subject of a sentence, the verb, and an object to form “she drinks coffee” or “he plays violin”)
3. Multiple choice exercises (oral selection from a list of different items)
4. Pronunciation of minimal pairs: *dip* vs. *deep*

These optional extensions could be incorporated:

5. Answering simple questions (e.g., “what is the girl drinking” with a picture showing that “she is drinking soda”)
6. Producing simple scene descriptions (e.g., “there is a house with a swimming pool”)

Preliminary estimates set the active target vocabulary to between 100 and 200 words and a perplexity of 5 to 10. However the final decision will be taken after the systematic evaluation of the user requirements.

The system evaluates the quality of the speech input and derives **feedback information** with respect to

1. Position of a mispronunciation
2. Kind of mispronunciation (phone quality, word stress patterns)
3. Possibilities and directions for improvement

The prototype system will be available in two versions

1. English pronunciation for Italian learners
2. English pronunciation for German learners

It will be integrated into existing and self-contained language learning courseware.

## 8 Feedback

Effective feedback involves, among other things:

- Users should be able to hear their own speech in comparison with the target and to be given guidance in analysing the difference. (This could also be done by means of wave forms, in which case they should be accurately aligned, although based on the user-surveys this would not be a particularly helpful mode of feedback.) The system should direct users to the significant differences and as far as possible inform them of the type of mismatch, especially the length and quality of individual phonemes.

- A context-sensitive help system could provide additional detailed guidance, perhaps with examples of correct articulation, stress pattern etc. The initial ISLE demonstrator will implement such contextual menus; the string uttered by the student will be shown on screen, with mispronounced words or areas clearly marked. A contextual menu will be attached to each word, allowing the user to choose from a variety of kinds of feedback.

For phone-type errors, the students will be able to:

- Hear the word (or a region containing the word) as they said it (it should be noted that this is a novel feature, as other similar products usually allow the student only to hear whole utterances)
- Compare it to the correct pronunciation
- Receive feedback telling them what they did wrong
- See tips on how to improve their production of that phone
- Jump to exercises specific to that phone

For stress errors, the students will be able to:

- Hear the word as they said it
- Compare it to the correct pronunciation
- Receive feedback telling them what they did wrong

The users' attempts at imitating the target sounds should be stored during a session and evaluated so that they can be reviewed at any time to compare levels of success.

### **9 Corpus-based Verification and Validation**

As suggested in the Requirements section, we will have large-scale off-line verification, using a test corpus collected for the ISLE project; and limited on-line verification or user trials. The on-line testing will focus at least as much on usability-type issues as on actual performance. This is likely to be more important in promoting and selling the system to English language teachers and learners; and in any case there will not be enough time, money, users, or exercises to implement a large enough system that would show statistically significant empirical results.

The on-line testing will be stand-alone (not tied to courseware) and will be completed by Leeds (with teachers) and Klett and UMilan (with students). It would be wonderful to have pretest-posttest differences to show, but due to time and finance constraints this will be possible, if at all, only for a very narrow range of problems. In addition to trying to measure performance increases for pronunciations, questionnaires will be designed for students and for teachers, to try to quantify their views of the software.

EAGLES and other Language Engineering projects have advocated user-centred evaluation (Maegaard 1998), which involves attempting to objectively assess how well the final system matches the user requirements specification. In addition to the above on-line testing, we will revisit this Report in a user-centred evaluation exercise.

For verification an annotated corpus of non-native speech will be collected, as we have not found a suitable existing corpus resource. The data collection specification includes the content, design and choice of recording texts (referred to as *prompts*); and the technical specifications of the speech recordings and the data collection software. The data annotation specification includes the data annotation scheme and annotation software.

The non-native speech corpus will be used to optimize the recognition and adaptation parameters for non-native speech and low-perplexity recognition tasks, and to evaluate the diagnosis of mispronunciations expected from intermediate learners of English. The corpus will therefore contain a representative sample of the target non-native accents and exercise types to be found in the final ISLE system. Two main sets of data will be collected from each speaker:

1. The *adaptation data* will be used to produce speaker-adapted non-native models for use in recognition experiments on the test data. The text prompts for the adaptation data recordings will also serve as the enrolment texts in the ISLE demonstrator. This adaptation data will allow us to evaluate how much enrolment data should be collected from each new ISLE user in order to give adequate non-native recognition performance. It will also allow the adaptation parameters to be optimized for the system.
2. The *test data* will be a series of short utterances for which low-perplexity syntaxes can be created. This will allow the recognition and diagnosis modules to be evaluated tasks equivalent to those used in the ISLE demonstrator system.

Data will be collected from 25 to 30 non-native, adult, intermediate learners of English, to include ten Germans and ten Italians. In addition, data from at least two native English speakers will be collected for test calibration purposes. The speaker set will be balanced for gender, age and accent variation as much as possible.

The data will be recorded using a tool developed at Entropic for the purpose of recording waveforms from a list of text prompts. The tool is able to load any list of prompts, and gives the user functionality to record, playback and view each utterance. The tool runs on both NT and Windows 95 and stores waveforms as WAV format files.

The error localization and mispronunciation diagnosis modules of the ISLE system will need to pinpoint errors at the phone level. In order to evaluate the performance of these modules, each utterance in the *test data* set will be annotated at the phone level. The final annotation will contain a transcription of how the utterance was spoken by the speaker in relation to a reference transcription containing a canonical native pronunciation. The phone-level reference transcription for each utterance will be produced automatically using Entropic's UK English recognizer running in a forced-alignment mode. *Adaptation data* will be only be verified at the word level but will *not* be annotated at the phone level.

Target words in the word-stress subset of the test data will be annotated with their expected stress pattern. The stress patterns are defined as sequences primary and secondary stress. The stress level will be annotated in the reference transcription alongside the vowels of the target word.

The phonetic annotations will be marked for three kinds of pronunciation errors at the phone level: substitutions, insertions and deletions, plus stress substitution errors. The error annotations will take the form E\_O where O is the expected form seen in the reference transcription and E is the observed form.

If time allows, it would be very useful to collect goodness of pronunciation scores from the human annotators at the utterance and word level and possibly at the phone level. This would give some finer indication of how well the subject is speaking and could be used to calibrate

and compare the localization and diagnosis components. Scores could be assigned on a scale from 1 to 4, where 1 is a native-like pronunciation and 4 is a very poor accentuated pronunciation.

## **10 Conclusions**

This report summarises Workpackage 1, covering ISLE user requirements and proposed solutions:

- The **requirements** of prospective users of ISLE are analysed in two ways: by consulting users, that is, English language teachers and learners; and through a survey of the market and competitor products. This requirements analysis provides ISLE partners with guidance on what users need in a computer system for pronunciation tutoring in EFL. The main tool in this exercise has been a variety of data-gathering exercises (questionnaire surveys, software evaluation and literature survey), with interpretation and analysis. The main conclusions relate to the model of spoken English the system should use, the kinds of learning activity it should provide and the type of feedback it should aim to give. We are further guided by an analysis of the market for educational software in general and ISLE-like systems in particular, including a survey of competitors and a detailed study of an example competitor product. We conclude that the assumptions at the start of the ISLE project are still valid, that is that plenty of CALL products are available but they provide inadequate feedback, and are not tailored to specific faults made by learners of specific nationalities.
- The **solutions** proposed by the ISLE project are outlined above. The ISLE system differs from previous systems because it (a) takes advantage of state-of-the-art recognition technologies that were obviously unavailable previously and (b) adds diagnosis and feedback at a much more specific level. The resulting system will be of value to ELT, and will fill a gap in the market having no direct competitors at the moment. Of course, active competitors are also involved in systems development; so it is critical to the success of ISLE to quickly implement the ISLE requirements specification into real-life software packages as demonstrators, and to license the ISLE SDK to third party publishers in order to quickly reach a “critical mass” for successful market penetration.

## **Appendix 1: Survey of language-learning products with speech recognition**

### **I. Auralog's *Aura-Lang*, *Tell Me More* and *Talk To Me* Product Series**

<http://www.auralog.fr/TTMegb.html>; <http://www.netg.co.uk/german/Tellme.htm>;  
<http://www.cornelsen.de>

Aura-Log was the first in Europe to incorporate speech recognition (SR) technologies in language learning software. These products are distributed all over Europe. In Germany they are distributed to the Institutional Market with the attribute "Professional" by NETG and to the Home Market by Cornelsen

#### ***TeLL me More***

Beginner, intermediate and advanced levels available (also "business" level in English)

Pronunciation screen. Over 100 hours of language work and over 1 000 exercises on each CD-ROM,

Key features: Videos, simple games, glossary, record/playback capability. A network version is available for educational institutions, training centres and companies.

Developer's product description: "*Tell Me More* is a "complete" method for learning a foreign language, covering all the essential areas: speaking, writing, comprehension, pronunciation and grammar. These areas are practised in depth, and the user has a high degree of control over difficulty levels."

Prices per language: DM 1,500.00 (single licence Professional edition)

DM 5,000.00 (licence for 10 Professional edition)

DM 99.00 (single licence, Private edition)

#### ***TaLk to Me***

Conversation Trainer available for English (beginner-confirmed), German (beginner, intermediate), Spanish (beginner, intermediate), French (beginner, intermediate), Italian (beginner, intermediate). Including listening, speaking and grammar exercises.

Key features: Pronunciation screen, games, record/playback capability.

Developer's product description: "*Talk to Me* is a 'conversational' approach to language-learning and focuses primarily on speaking and listening."

### **II. *DynEd New Dynamic English*, <http://www.dyned.com/dyned/eng/denmain.htm>**

Suitable for both beginners and advanced learners; listening, speaking and reading exercises, ca. 200 hours in total. Spanish, Japanese, Korean, Italian, Portuguese, French and English-only versions available.

Key features: Speech recognition, interactive video lessons, difficulty adjusted, keeps track of performance, effective sequencing results in long-term learning, on-line glossary, record/playback capability.

### **III. Langenscheidt group's language trainers (*English in 30 Days*)**

<http://www.aspect.de/doc/forlang.htm>; <http://www.langenscheidt.de>

Language learning software using Aspect's speech verification software (see below) are available under the name of Langenscheidt, Humboldt-Verlag and Hexaglot. Humboldt-Verlag's language learning products (e.g. "Französisch in 30 Tagen") are available in Germany for the languages English, French, Italian, and Spanish. Price 50 ECU

#### **IV. Digital Publishing *Interactive Journey through Language, The 20th Century***

<http://www.digitalpublishing.de/isr.htm>;

[http://www.speech.be.philips.com:100/bin/owa/psp\\_s\\_press?xid=747](http://www.speech.be.philips.com:100/bin/owa/psp_s_press?xid=747)

Based on a co-operation with Philips Speech Processing and the use of *FreeSpeech 98*.

Available for English, French, Spanish and Italian.

Key features: Speech recognition *IntelliSpeech*, realistic situations, video moderated, pronunciation exercises, integrated dictionary, several hours of photo stories, intelligent error analysis, "Smart Pitch Control" (audio speed adjustment), simultaneous translation.

Digital Publishing claims that *IntelliSpeech* is the world's first speech recognition software developed exclusively for language learning software in contrast to SR that is developed to control machines/applications. This must be qualified because Auralog is working in the same direction of course and *IntelliSpeech* is based on the Philips engine (*FreeSpeech98*), the origin of which is SR for voice control applications, too.

*IntelliSpeech* is supposed to give intelligent feedback on the quality of speech and tries to find out how close a learner can get to a model speaker speaking in mother tongue. For this, each word in the dictionary is available with a reference spoken by a male and female as well as an old and a young speaker. After a first look-up in the phoneme database the learner's speech input is compared in several stages to the stored references in order to provide differentiated feedback.

Developer's product description: "*FreeSpeech98* allows the user of a language learning program to test his pronunciation interactively," according to Armin Hopp, Managing Director of Digital Publishing. "Thanks to the automatic speech recognition feature, he can now immediately check if he has got the words right."

#### **The Learning Company, *Learn to Speak English* <http://store.learningco.com>**

30 chapters based on real-life situations prepare families for an unlimited number of everyday situations.

Key features: Extensive audio and video of a variety of native speakers builds comprehension of the language as it's spoken. Advanced speech-recognition (Windows version only) and record/playback technologies evaluate and improve pronunciation. 400-page text/workbook with extended grammar exercises provides additional in-depth study. Over one hour of original QuickTime® movies presents the native culture. Listening and grammar exercises.

Platforms Supported: Windows® 3.1 and higher; Macintosh® System 7.0.1 and higher  
Price US\$ 80 (In USA), ECU 100 in Germany (as *Sprechen wir Englisch* by TLC Tewi)

#### **Syracuse Language System/Living Language Multimedia/Random House: *Start speaking a new language today!* (formerly *TriplePlay Plus!*) and *Lets Talk*, <http://www.syrlang.com/>**

##### ***Start speaking a new language today***

Windows® CD-ROM, age 12-adult, versions available: Chinese English French German Hebrew Italian Japanese

Developer's description: "Fun and rewarding - no rules, no drills! Start speaking your new language right away, in multimedia learning activities. Begin by perfecting your pronunciation of key words with an interactive trainer. As you progress through three skill levels, master words and phrases - even conversations! Speech recognition provides rapid feedback to reward correct answers and pronunciation. Record/playback compares you with native speakers.

Price US\$50

***Let's Talk***

Vocabulary/pronunciation skill builder, Windows® CD-ROM, versions available: Spanish, French, German, Italian, English.

Developer's description: "The easy way to learn vocabulary, improve pronunciation, and build fluency, Let's Talk teaches over 2,200 words in Spanish, French, German or Italian. Native speakers demonstrate correct pronunciation, and full-color photographs illustrate each word and phrase. Learning games cover more than 100 topics. Speech recognition and record/playback provide prompt feedback.

Let's Talk English teaches over 2,500 English words and phrases to speakers of Chinese, French, German, Italian, Japanese, Korean, Portuguese and Spanish."

Price US\$30

***Appendix 2: Online References***

BECTa - British Educational Communications and Technology Agency (formerly NCET - National Council for Educational Technology): <http://www.becta.org.uk>

CTI Centre for Modern Languages Online Software Database:  
<http://www.hull.ac.uk/cti/searchdb.htm>)

Educational Multimedia Task Force: Progress Report as of Second Quarter of 1998,  
<http://www2.echo.lu/emtf/en/rpt0798english.html> EUROCALL - European Association for Computer Assisted Language Learning: <http://www.hull.ac.uk/cti/eurocall.htm>

ESOL - English as a Second Language <http://www.becta.org.uk/resources/esol/>

IATEFL International Association of Teachers of English as a Foreign Language:  
<http://www.iatefl.org/>

iBusiness - Multimedia News and Trends: Interactive Business NET im Hightext Verlag:  
<http://www.ibusiness.de/>

RECALL website <http://www.infj.ulst.ac.uk:80/~recall/index.html>

The British Council: <http://www.britcoun.org/>

Verzeichnis lieferbarer elektronischer Medien (VLE): <http://www.buchhandel.de>