# ISLE

## Validation Report

Project: LE4-8353
Deliverable: D5.1

| | |
|---|---|
| Version | 5 |
| Date | 03.07.2000 |

# ISLE Deliverable

| | |
|---|---|
| **Project Number** | **LE4-8353** |
| **Project Title** | **Interactive Spoken Language Education [ISLE]** |
| **Deliverable Type** | **Tool, Report** |
| **Distribution** | **Restricted** |
| **Deliverable ID** | **D5.1** |
| **Expected Delivery Date** | |
| **Actual Delivery Date** | **06 June 2000** |
| **Title of Deliverable** | **Validation Report** |
| **Authors** | **ULeeds [Howarth], Umilan [Pezzotta, Galbiati, Bisiani]** |

| OT | RE | SP | PR | TO |
|---|---|---|---|---|
| Other | Report | Specification | Prototype | Tool |

| C | P | R |
|---|---|---|
| Consortium | Public | Restricted |

Revision History

| Version | Date | Status | Author(s) |
|---|---|---|---|
| 1 | 03-09-1999 | Draft | U Milan [Pezzotta, Galbiati, Bisiani] |
| 2 | 06-04-2000 | Final Part I | U Milan [Pezzotta, Galbiati, Bisiani] |
| 3 | 15-05-2000 | Draft Part II | U Leeds [Howarth] |
| 4 | 06-06-2000 | Draft | edited by Menzel |
| 5 | 1/7/2000 | Final | R. Bisiani |

# Part I: Executive summary

The goal of the ISLE project is to build a tool to help adult intermediate learners of English improve their pronunciation, using speech recognition technology.

This report describes:

- The methodology and the results of an on-line evaluation carried out by asking foreign language students and teachers to report their experiences when using the ISLE demonstration system,
- the validation tools, a set of software applications built to analyze the performance of the diagnosis components of the ISLE demonstrator;
- the overall results of on off-line evaluation of the diagnosis components.

Specifications for the data collection are provided in ISLE report D31.

Detailed performance results can be found in:

- ISLE report D4.1.: Error localization component;
- ISLE report D4.2.: Detection of mispronunciations in non-native speech;
- ISLE report D4.3.: Word-stress detection

In particular, this report details:

- The set-up of the on-line evaluation (the subjects included, the procedure and the questionnaire),
- a statistical summary of the judgments by students and teachers and a compilation of individual observations and remarks,
- the data structures used by the validation tool,
- the analysis performed by the tool to validate the diagnostic results for phone and stress against the human annotation of a non-native language speech corpus,
- the experiments to find the best thresholds for recognition and localization to be used in the ISLE demonstrator

This is a public report.

# Part II: The on-line evaluation

## 1. Trialling

For the purposes of testing the effectiveness of the ISLE demonstrator, the system was trialled with groups of adult non-native speakers of English from Italy and Germany, non-native teachers from Germany and native-speaker teachers in the UK, a total of 28 subjects:

| | |
|---|---|
| *University of Milan, Bicocca* | 6 Italian-speaking learners |
| *Klett verlag, Stuttgart* | 9 German-speaking learners |
| | 8 German-speaking teachers |
| *University of Leeds* | 5 English-speaking teachers |

## 2. Procedure

The demonstrator was installed at each location from CD and tested out by the project partners. In each case an evaluator was identified, who would supervise the trialling sessions. Instructions were distributed to the supervisors (see Appendix 1) and an introduction to the ISLE project was given to each volunteer (Appendix 2).

## 3. Data collection

Two sources of information were used for data collection:

- a questionnaire for the volunteer users
    - a version for learners, and
    - a slightly different version for teachers
- a report form for evaluators to note any problems experienced or reactions by users (Appendix 3)

## 4. Data analysis

There follow the collated data from the various sources.

### 4.1. Native English-speaking teachers' questionnaires

Not all the separate comments are recorded here where there is considerable overlap.

**1. feedback**

**1.1 Is the feedback easy for a learner to understand?**

| v.easy | easy | neither | difficult | v.difficult |
|---|---|---|---|---|
| | **4** | **1** | | |

Comments:

"Explanation and the chance to listen again to the native speaker as often as the student wishes is good"

"Easy but often inaccurate or vague"

"Clear but limited in scope. No suggestions are given"

**1.2 Do you feel the feedback would be accurate in identifying their errors?**

| v.accurate | accurate | neither | inaccurate | v. inaccurate |
|---|---|---|---|---|
| | **1** | | **3** | **1** |

Comments:
"One feedback comment confused the computer 'model' pronunciation with the speaker's pronunciation"
"I don't think it's very clear in identifying whether the error is one of stress or of sound production"

**1.3 Did you mispronounce sounds that the program didn't identify?** yes **4** no **1**

Comments/examples:
"It doesn't recognise other variants"
"bratwurst"
"It picked up few consonant errors (eg d and p)"

**1.4 Did the program falsely identify errors?** yes **5** no

comments/examples:
"Deliberate mispronunciations of /EY/ were almost never picked up"
"As a native speaker, the program constantly corrected my pronunciation, which, I suppose, is RP! A little worrying."
"This is rather difficult to prove, but the speaker's pronunciation of /EH/ in one waord was understood as /AE/ by the computer."
"eg in 'wonderful' /UH/ for /AH/"

**1.5 Would the feedback help learners to improve their pronunciation?**

| v.well | well | neither | badly | v. badly |
|---|---|---|---|---|
| **1** | **1** | **2** | **1** | |

Comments:
"I feel it depends entirely on a particular example. The speaker's pronunciation of 'glass' was very similar to the computer's, but was seen as a 'problem'. When repeated 5 times in 5 different ways the screen comment was 'good try'"
"For some 'problem' words there was no concrete model to listen to, only advice to keep practising. Not very helpful."
"It seems to be useful for specific sentences and sounds"
"Identifying the error, explaining it, isolating the sound, repeating it has to be useful"
"Often but not always a mispronounced word was correctly identified, but the diagnosis was either vague, or focused on the wrong syllable, or on a vowel instead of a consonant."

**2. material**

**2.1 Is the language users have to speak realistic?**

| v. realistic | realistic | neither | unrealistic | v. unrealistic |
|---|---|---|---|---|
| | **3** | **1** | **1** | |

Comments:
"The language itself is realistic but the delivery is unconvincing in the dialogues. The speakers sound bored."
"Going on a barge trip isn't very common".
"It seemed relatively realistic, although I'm not sure I would say 'I'll have a pizza and a soda to drink'"

**2.2 Are the instructions clear?**

| v. clear | clear | neither | unclear | v. unclear |
|---|---|---|---|---|
| | **1** | **2** | **1** | **1** |

Comments:
"They vary. Many of the buttons are hard to find, in odd places or confusingly named ('micrtophone' can be either the microphone or the head in profile"
"Initially the first interface is very confusing. The other screens are a bit confusing"
"It does depend on how well the listener understands the symbols, which are not always clear eg the side panel"
"Extremely difficult to follow. Not obvious."
"Many things missing. The initial 'Start Program' button was above the general blurb, which you had to read first. Needs much more careful thought"

**3. design**

**3.1 Are the exercises/activities interesting?**

| v.interesting | interesting | neither | uninteresting | v.uninteresting |
|---|---|---|---|---|
| | **1** | **2** | **1** | **1** |

Comments:
"Repetitive but I suppose this is hard to avoid"
"They are neutral- standard and bland- but clear enough."
"I don't think the interest level is very high as the situational dialogues are very conventional."
"How 'interesting' can you expect an exercise to be?"
"Limited exercise types. Not particularly stimulating sentences"

**3.2 Is the program visually attractive?**

| v.attractive | attractive | neither | unattractive | v.unattractive |
|---|---|---|---|---|
| | **4** | | **1** | |

Comments:
"The colours are repetitive"
"Quite easy on the eye and friendly if a little esoteric."
"I can't see the rationale behind the opening web page design"
"Stylish to look at, but isn't always easy to find the buttons to click"
"Attractive but I found it hard to follow logically. I prefer a linear design, rather than a globular flowing design. The colours are a bit insipid"

**3.3 Is the language varied enough?** yes **1** no **2** don't know **2**

Comments:
"There seems to be a model and variants away from this may be considered incorrect"
"If it's only dialogues, then obviously it isn't varied enough for pronunciation purposes"

**3.4 Would you recommend your learners to use the program (again / more than once)?** yes **2** no **3**

Comments:
"Yes, because repetition plus clear exemplification is important with pronunciation. Students can go over each item as often as they like"
"Yes, if they were having problems with individual sounds"
"No, nowhere near accurate enough"
"Not yet, needs much more work"
"No, it is too unreliable for students to work with on their own"

**4. learning**

**4.1 Does the program cover the most important pronunciation features?** yes **3** no don't know **2**

Comments:
"Yes, stress, minimal pairs"

**4.2 What is missing?**

Comments/examples:
"Links between words"
"weak forms versus full forms"
"Rhythm patterns, linking"
"Intonation patterns"

**4.3 Is the target pronunciation appropriate?** yes **5** no

Comments:
"It's the one most students want"
"I heard 2 different accents. How does the computer feedback differentiate between 'errors' and variants? What is the standard"
"Yes, however scope for variations wold be useful"

**4.4 Is the practice at the right level for intermediate learners?** yes **4** not sure **1**
Comments:
"A bit easy?"

**4.5 Would this material contribute to the development of your students' spoken English?** yes **3** no **2**
Comments:
"Yes, if it were seen as a resource with certain limitations"
"No, too many variables and unsolved problems"

**4.6 What additional features would improve the program?**
Comments:
"A link between 'teacher's' demo, pronunciation of a phrase and the diagnosis of a problem would be essential- occasionally the 'teacher' seemed to be making the same 'mistake' as the student e.g. using the weak form of 'than'"
"Using the phonetic alphabet instead of highlighting a letter in the word e.g. PROBLEM: an Italian speaker might think it should be pronounced /with /EH/ not /AX/."
"Ability for students to enter his/her own speech (language examples) and have feedback"
"Clearer symbols/buttons- perhaps fewer"
"Easier initial stage: reading individual sentences for the computer to adjust to the speaker's accent is tedious and the sentences are disjointed when separated by time gaps"

**4.1.1 Evaluator's reports**

**1. Problems experienced by the user:**
"Constant problems with the volume level"
"The position of the start button at the top of the opening page is confusing"
"What to do now? Frequent pauses to find out what to do next"
"No warning that sound is going to start"
"One user didn't discover the text of the dialogue"
"Took a long time to discover the exercises"
"Not easy to move smoothly between the dialogues and the exercises"

**2. Reactions of the user (visual or verbal):**
"The adaptation is tedious. Why so long?"
"Which is the settings button? What is the difference between the settings? Not clear what the arrow buttons do"
"The colour of the buttons means they don't stand out"
"Tried using an Italian accent in the adaptation phase"
"Why are text exercises included?"
"Slow response caused multiple clicking"
"A lot of clicking at the wrong time"
"Odd design"
"Liked the layout and background; user-friendly appearance"
"not clear what to do after listening to the dialogue"

**3. Components of the program used:**
All users used all available components.

**4. Other comments:**
"It was an exercise in discovery. Without the evaluator's help it takes a very long time to find out how to use the program"
"A high level of strictness results in a lot of non-errors"
"Very hard to hear difference between teacher and student versions"
"Does 'than' in the system's dictionary have a weak variant?
"Very unreliable/inaccurate performance doesn't inspire confidence"
"'In the Office' dialogue: not possible to have text and listen to dialogue at the same time"
"The length of the adaptation phase is disproportionate to the amount of practice material available"
"Necessary to go back to the text (not easy) to answer the questions"

## 4.2 German-speaking teachers' questionnaires

**1. feedback**
**1.1 Is the feedback easy for a learner to understand?**

| v.easy | easy | neither | difficult | v.difficult |
|---|---|---|---|---|
| | **4** | **3** | **1** | |

**1.2 Do you feel the feedback would be accurate in identifying their errors?**

| v.accurate | accurate | neither | inaccurate | v. inaccurate |
|---|---|---|---|---|
| | **5** | **3** | | |

**1.3 Did you mispronounce sounds that the program didn't identify?** yes **4** no **3**

**1.4 Did the program falsely identify errors?** yes **5** no **3**
comments/examples:

**1.5 Would the feedback help learners to improve their pronunciation?**

| v.well | well | neither | badly | v. badly |
|---|---|---|---|---|
| | **8** | | | |

**2. material**
**2.1 Is the language users have to speak realistic?**

| v. realistic | realistic | neither | unrealistic | v. unrealistic |
|---|---|---|---|---|
| | **6** | **2** | | |

**2.2 Are the instructions clear?**

| v. clear | clear | neither | unclear | v. unclear |
|---|---|---|---|---|
| **2** | **6** | | | |

**3. design**
**3.1 Are the exercises/activities interesting?**

| v.interesting | interesting | neither | uninteresting | v.uninteresting |
|---|---|---|---|---|
| **1** | **6** | **1** | | |

**3.2 Is the program visually attractive?**

| v.attractive | attractive | neither | unattractive | v.unattractive |
|---|---|---|---|---|
| | **5** | **2** | **1** | |

**3.3 Is the language varied enough?** yes **6** no **1** don't know **1**

**3.4 Would you recommend your learners to use the program (again / more than once)?** yes **8** no

**4. learning**
**4.1 Does the program cover the most important pronunciation features?** yes **5** no don't know **3**

**4.3 Is the target pronunciation appropriate?** yes **5** don't know **3**

**4.4 Is the practice at the right level for intermediate learners?** yes **7** not sure **1**

**4.5 Would this material contribute to the development of your students' spoken English?** yes **5** don't know **3**

**4.6 What additional features would improve the program?**
Comments:
"More female voices"
"Analysis of free spoken text production"

## 4.3. Users' questionnaires: Italian learners

**1. feedback**
**1.1 Is the feedback easy to understand?**

| v.easy | easy | neither | difficult | v.difficult |
|---|---|---|---|---|
| **2** | **3** | **1** | | |

**1.2 Do you feel the feedback is accurate in identifying your errors?**

| v.accurate | accurate | neither | inaccurate | v. inaccurate |
|---|---|---|---|---|
| **3** | **3** | | | |

Comments:
"I like very much the feedback in which I can understand the way I have to pronounce a phone, reading another word with the same phone"

**1.3 Did you make errors that the program didn't identify?** yes no **6**

**1.4 Did the program identify errors you thought were correct?** yes **3** no **3**
Examples:
'business'
'photographer'

**1.5 Does the feedback help you to improve your pronunciation?**

| v.well | well | neither | badly | v. badly |
|---|---|---|---|---|
| **2** | **4** | | | |

Comments:
"To me it is very useful the fact that it is possible to see the errors I made and to repeat similar words (IMPROVE window) to learn the pronunciation of a single phone"
"I think that 'IMPROVE' and 'Practice the phones you have the most problems with' are really great ideas"

**2. material**
**2.1 Is the language you have to speak realistic?**

| v. realistic | realistic | neither | unrealistic | v. unrealistic |
|---|---|---|---|---|
| | **6** | | | |

Comments:
"In the demo I saw Lesson 5 and I think that it describes a realistic situation"

**2.2 Are the instructions clear?**

| v. clear | clear | neither | unclear | v. unclear |
|---|---|---|---|---|
| | **6** | | | |

Comments:
"I have only a consideration to do: In "Oral Exercise" there is an instruction that tells: "Please click on the microphone and read the sentence in the box" but on the window there are two "Microphone" buttons, and this can do muddle. "
"In Standard exercises there is not legend that tell me the meaning of the GREEN/RED feedback."

**3. design**

**3.1 Are the exercises/activities interesting?**

| v.interesting | interesting | neither | uninteresting | v.uninteresting |
|---|---|---|---|---|
| **1** | **4** | **1** | | |

Comments/examples:

"In particular I like very much the fact that I can run different kinds of exercises….this make the demo not boring."

" To me it is very boring the Standard Exercise TRUE/FALSE."

" The exercise I prefer is "Listen and Repeat"."

**3.2 Is the program visually attractive?**

| v.attractive | attractive | neither | unattractive | v.unattractive |
|---|---|---|---|---|
| **4** | **2** | | | |

Comments:

"I like very much the menu to choose the exercise"

"The interface is very beautiful"

" I like very much the window "Arrival to Manchester", in which I can select the lessons."

"I like very much the colours defined for the windows (the background colours): they are very relaxing."

**3.3 Is the language varied enough?** yes **6** no

Comments:

" I think that there will be no tool enough detailed so that it can be considered enough to learn English perfectly. But, to me it is great the function associated with the "ABC" button, through which I can pronounce a lot of words associated with a particular phone."

**3.4 Would you use the program again / more than once?** yes **6** no

Comments:

" For my English it would be a good thing to use this demo again. In particular I like FREE CHOICE Oral exercise."

" For my English it would be a good thing to use this demo again."

" It is a good idea the fact that I can see and read the dialogue. Besides it is great the way the lesson is introduced."

" I like very much this demo, in particular the fact the it is able to correct my pronunciation at phone level."

" I like it very much: I saw other tools and I think this is the best!!!"

**4. learning**

**4.1 Does the program cover the most important pronunciation features?** yes **4** no

[no answers were given to 4.2 and 4.3]

**4.4 Is the practice at the right level for you?** yes **5** no **1**

Comments:

" No, because my English is really bad, and so the level is really high to me."

" My English is not so good and this test confirms it. So I think the level is right"

**4.5 What additional features would improve the program?**
Comments/examples:
"ToolTips for the buttons would be very useful."
"Change the Navigational arrows to make them more intuitive: for example the arrows to change the exercise (the external ones) can be vertical. The background of the windows is too homogeneous: the windows would be more attractive putting more colours on it (changing for example changing the colour of the windows' frame)."
"Put a legend to the STANDARD exercises to tell the user the meaning of the GREEN/RED feedback.
Define a button "Give me the Correct Translation " in the Translate Exercises.
"The buttons on the windows seem image and not buttons: it would be better that when the user pushes on a button, it changes its appearance."
" Add a progress–bar in READ-REPEAT exercise to tell me the velocity through which I have to pronounce the phrase.
Define a new type of exercise as mix of "READ-REPEAT " and "LISTEN-REPEAT", in which I can hear the utterance but I can have the text of the phrase to help to repeat."

**4.3.1 Evaluator's reports**

**1. Problems experienced by the user:**
"There are some problems about the voice-feedback for the diagnose errors. For example in the "Read and Repeat" exercise with the phrase "They asked if I wanted to come along on the barge trip" the speaker make an error on the word "ASKED" and when I pushed the "teacher" option on the pop-menu' the demo "said" "THEY ASKED" and not only "ASKED"."
" In some occasions the "Teacher" button doesn't run."
" In the dialogue the Mr. Rossetti's voice is low and so difficult to understand.
To do ORAL exercises why does the user must listen for the dialogue?
Buttons aren't intuitive."
" In a occasion in the "Fill in the blank" exercise the inserted word covers the fixed text.
"Fill in the blank" option in the menu' doesn't work: the user access this exercise only through the navigational arrows."
" In "Standard Exercise"
The exercise "Translate" doesn't run;
The external navigation arrow doesn't run
In "TRUE/FALSE" exercise there is written:
"After listening the dialogue, please answer these questions with YES or NO "
But if the user hasn't listened for the dialogue before, there is no way to listen for it."
" In some occasions the student had difficult to understand how to come to the previous window and, in general, to capture the meaning of the buttons."

**2. Reactions of the user (visual or verbal):**
"My impression was that the speaker seemed to be very enthusiastic…he told me that the demo's interface are beautiful, and in particular he liked very much the diagnose-feedback of the error (possibility to listen for the correct pronunciation and to practice on a wrong word [IMPROVE window])"
" She told me that the demo is very interesting, but she was very perplexed, because the meaning of the buttons often it is not clear."
" The speakers told to me that the ORAL exercises are great, but he was very boring to do the STANDARD exercises."
"The user seemed to be enthusiastic about the demo's interface and, in particular, she liked the diagnose-feedback of the error."
" He had some trouble with the Oral Feedback in "Improve": he told me that it was often too fast to understand."
"She told me that the demo is very beautiful, but sometimes she was in difficulty to understand the meaning of the buttons.
Her precise words: "The buttons are intuitive for nothing"."

**3. Components of the program used:**
All Oral and standard exercises were used by all subjects.

**4. Other comments:**
"The speaker was very worried to test the demo, because he said that his english was not so good.
So he made a lot of errors only because he spoke on the microphone very slow to try to pronounce correctly the phrases….but the demo often didn't wait for him."
"This student speaks English very well: so she can be considered a good tester to understand if the demo really finds the right pronunciation errors."

## 4.4. Users' questionnaires: German learners

**1. feedback**
**1.1 Is the feedback easy to understand?**

| v.easy | easy | neither | difficult | v.difficult |
|---|---|---|---|---|
| **1** | **7** | **1** | | |

comments/examples:
" The feedback omitted the same pronunciations that were slightly different"

**1.2 Do you feel the feedback is accurate in identifying your errors?**

| v.accurate | accurate | neither | inaccurate | v. inaccurate |
|---|---|---|---|---|
| **1** | **3** | **5** | | |

comments/examples:
" Often the comment is inspecific, the problem is not explained"
" System failed to identify utterances which were quite different from each other"
"Sometimes surprising!"
" The program should stress the main mistakes"

**1.3 Did you make errors that the program didn't identify?** yes **5** no **4**
comments/examples:
" If you are reading the words of the word list (Improve) the program only identifies the error you are improving in that exercise, but neglects any other errors you make."

**1.4 Did the program identify errors you thought were correct?** yes **5** no **4**
comments/examples:
"pan, bag"
"Of course!"
" I couldn't detect any difference between the "teacher" and "student""

**1.5 Does the feedback help you to improve your pronunciation?**

| v.well | well | neither | badly | v. badly |
|---|---|---|---|---|
| | **9** | | | |

comments/examples:
"The possibility of listening to the examples in the diagnosis is helpful"
"It is a problem that I'm not quite fixed to either British or American English, so the pronunciation might be correct in another context, but is wrong here"
"Asked!"
"asked, wanted, address"

**2. material**
**2.1 Is the language you have to speak realistic?**

| v. realistic | realistic | neither | unrealistic | v. unrealistic |
|---|---|---|---|---|
| **1** | **8** | | | |

**2.2 Are the instructions clear?**

| v. clear | clear | neither | unclear | v. unclear |
|---|---|---|---|---|
| **2** | **4** | **2** | **1** | |

comments/examples:
"How the user is guided by the program could be better (e.g. If you click on a button, an explanation could be displayed). Clear buttons (without explanation are not useful!)"
"The spoken instructions are clear, the written ones are sometimes not (e.g. lesson 2, "Click on the microphone", which microphone? There are two of them"
"After accommodation to buttons and instructions, the handling was unclear"

**3. design**

**3.1 Are the exercises/activities interesting?**

| v.interesting | interesting | neither | uninteresting | v.uninteresting |
|---|---|---|---|---|
| **2** | **5** | **2** | | |

comments/examples:
"Because it is realistic"

**3.2 Is the program visually attractive?**

| v.attractive | attractive | neither | unattractive | v.unattractive |
|---|---|---|---|---|
| | **7** | **1** | **1** | |

comments/examples:
"Buttons do not use clearly understandable metaphors, different metaphors / symbols are used in different contexts of the program"
"I prefer clear lines, not this bubble-gum outfit"
"A simple interface that points out the main functions"

**3.3 Is the language varied enough?** yes **6** no **1** don't know **2**

**3.4 Would you use the program again / more than once?** yes **7** no **1**

comments/examples:
"for an advanced speaker / learner the program is too detailed. I would prefer to play in an English speaking country and adapt to what I hear."
"It would be better if there were more dialogues per unit offered, as often they are repeated"

**4. learning**

**4.1 Does the program cover the most important pronunciation features?** yes **7** don't know **2**

**4.2 What is missing?**

comments/examples:
"One word is often repeated one after another, therefore there is a lacking in the possibility to listening again"

**4.3 Is the target pronunciation appropriate?** yes **7** don't know **2**

**4.4 Is the practice at the right level for you?** yes **7** no **2**

comments/examples:
"Too difficult (in the middle sector)"
"I had not learnt some of the vocabulary"
"Sometimes too hard"
"The vocabulary could be harder as it is too easy"
"Could be a useful help to correct the pronunciation"

**4.5 What additional features would improve the program?**
comments/examples:
"Different speakers and speeds of the teacher"
"Translations / Dictionary in the background"
"The possibility of taking out short sequences from the whole sentence and then it works!!"
"more pictures, more interface features using the pc, the actual version is similar to a tape exercise"

**4.4.1 Evaluator's report**

**1. Problems experienced by the user:**
"Open the feedback was "I wouldn't understand""
"The texts for adaptation are too small - difficult to read"
"Difficulty to speak an unknown word"
"Sometimes it couldn't be defined where the problem is (no specific diagnose given)"
"Listen and repeat: If the sentence couldn't be understood it is difficult to repeat and got a qualified reaction!"
"User tried to adapt to the way of speaking of teacher (faster), but this is not accepted"
"Examples for wrong and correct pronunciation in diagnose stage often hard to understand"
"Adjusting the microphone not found"
"Once the introduction text didn't end, it took some trials to jump to the dialogue"
" Sometimes forgets to press the speak button, or presses it and doesn't speak."
" The pop-up menus are sometimes too small, it is difficult to hit them accurately with the mouse."
"Dialogue overall is not easy to understand"
"Lesson 2, "Click the microphone", could be understood as to click the oral exercises"
"To pronounce unknown words"
"Click-Speak co-ordination"
"Dialogue interface, not clear"
"Exit from sub-chapters"
"Directions for user have been ignored/overlooked"

**2. Reactions of the user (visual or verbal):**
"likes clicking"
"repeats the exercise many times to improve outout / result"
"uses seldom the diagnose function"
"Experienced that a click on a blue word (improve stage) will get the teacher to speak the word"
"Impression that the user speaks with less melody than the user and then it causes unspecified problems."
"Long sentences cause problems: If the user concentrates on specific problems and is to improve, a problem may arise at another place. Would it be possible to practice parts of a sentence?"
"Laughs when own voice is heard"
"Happy with success"
" Was happy with good results"
"After some experience, more often repeated the sentences than look for diagnose and do improvement"
"Suprised, how differentiated the reaction of the system is."
"Immediately changed the cursor of "How strict should I judge?", to a lower position"

**3. Components of the program used:**
All subjects did Lessons 1 and 2.

**4. Other comments:**
"Impatient clicking confuses the system and takes a long time to decide what should be the next step. But never crashes."
"Stopping the introduction and changing to the dialogue is not explained to the user. Also not good: if the button (read and listen) is pressed, the dialogue stops."
"Sometimes the words spoken as examples are to short and difficult too analyse"
"Now and then frustrating when many mistakes are in one sentence"
"Minimal pairs ( ae/ and /e/ - /ae/ seems to ask for an /a/ sound."
" Very often the diagnose couldn't give special hints"
" Problems with the system:
a). Free choice: sound of the sentence components became inactive
b). The teacher's voice was not active"
" Sometimes (apart from clicking open to "teacher") the spoken word was different to the one displayed as wrong"
""Listen and Repeat" exercises are difficult, because often the student doesn't understand correctly the content of the sentence."
"If the German speaker mimicks the teacher, and speaks as fast as he, the system doesn't understand. But the student doesn't get a hint what's the reason for this problem."
"Minimal pairs, it would be more comfortable for the user, if already the first word would appear in blue, so that the student knows which word to read and speak."
"The program was not quite correct: student said, "eightieth" instead of "eighteenth" and the feedback was the incorrect "th""
"Problem: do the exercise "Free Choice", if the student crosses one of the words, with the mouse, the text dissapprears (greys)"
"Difficulties because questions and answers don't fit correctly to the texts, (causes frustration to the user)"
"Part of a sentence which was spoken completely was analysed!"
""Build the sentence", in case the answer is not the correct one this should be mentioned, but nevertheless the pronunciation should be scored and corrected"
"Help was necessary, because interface is not clear (controls had been explained)"
"Speak-Click interaction needs practice/adjustment??

# Part III: The validation tool

The validation tool is written in Visual Basic 6.0 and stores all data in an Access 97 database for further analysis. The database permits to access to the data outside the Tool.

The tool's functions are:

- Fill the annotation table, using data from annotators;
- Fill the diagnose table, running IHAPI and the diagnostic routines;
- Fill the compare table, merging the two tables and aligning the results;
- Generate graphs for phone and stress analysis.

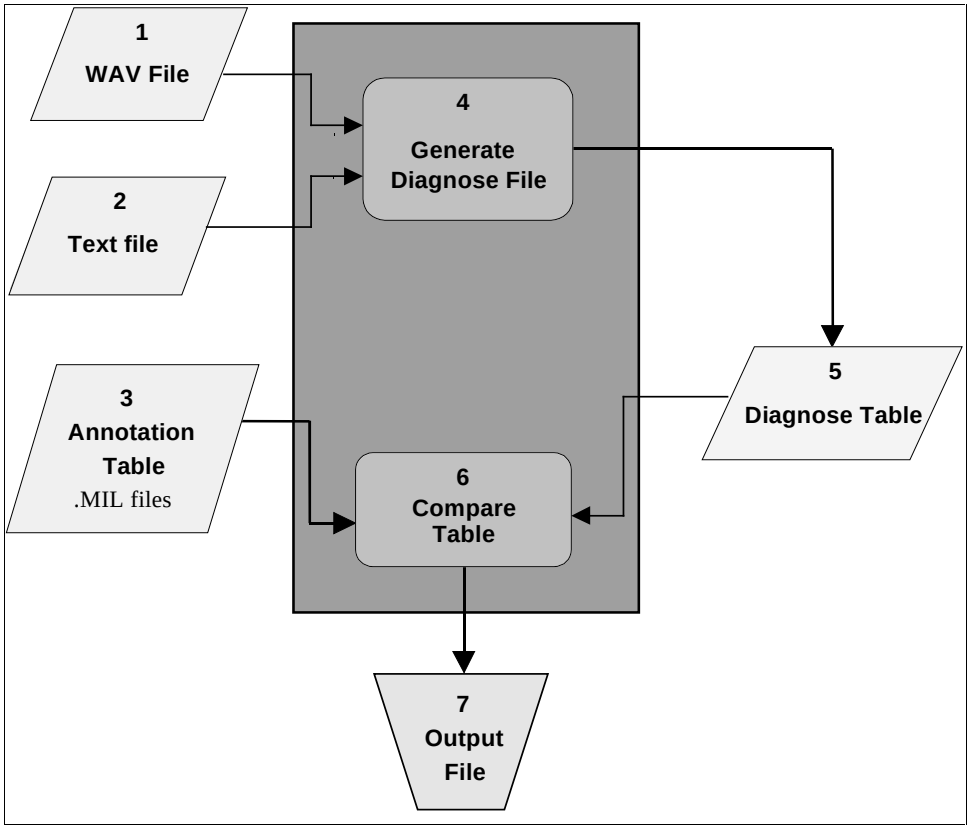


**Figure 1: The Structure**

Where:

- **1.** The **.wav** file associated with the utterance;
- **2.** The phrase's text to build the syntax;
- **3.** This table is filled through the **.MIL** files;
- **4.** The processing performed by the ISLE DLLs (the same that are used in the demonstrator);
- **5.** The table generated by the diagnostic routine;
- **6.** Comparison table between the annotation (**3**) and the diagnose (**5**) tables.
  There are two types of comparison: phone and stress comparison;
- **7.** The tool's output: it is represented through graphics in the **.xls** file.

# 4. The output of the validation tool

The validation tool can produce graphs for phone and stress error analysis.

## 4.1. Phone error analysis

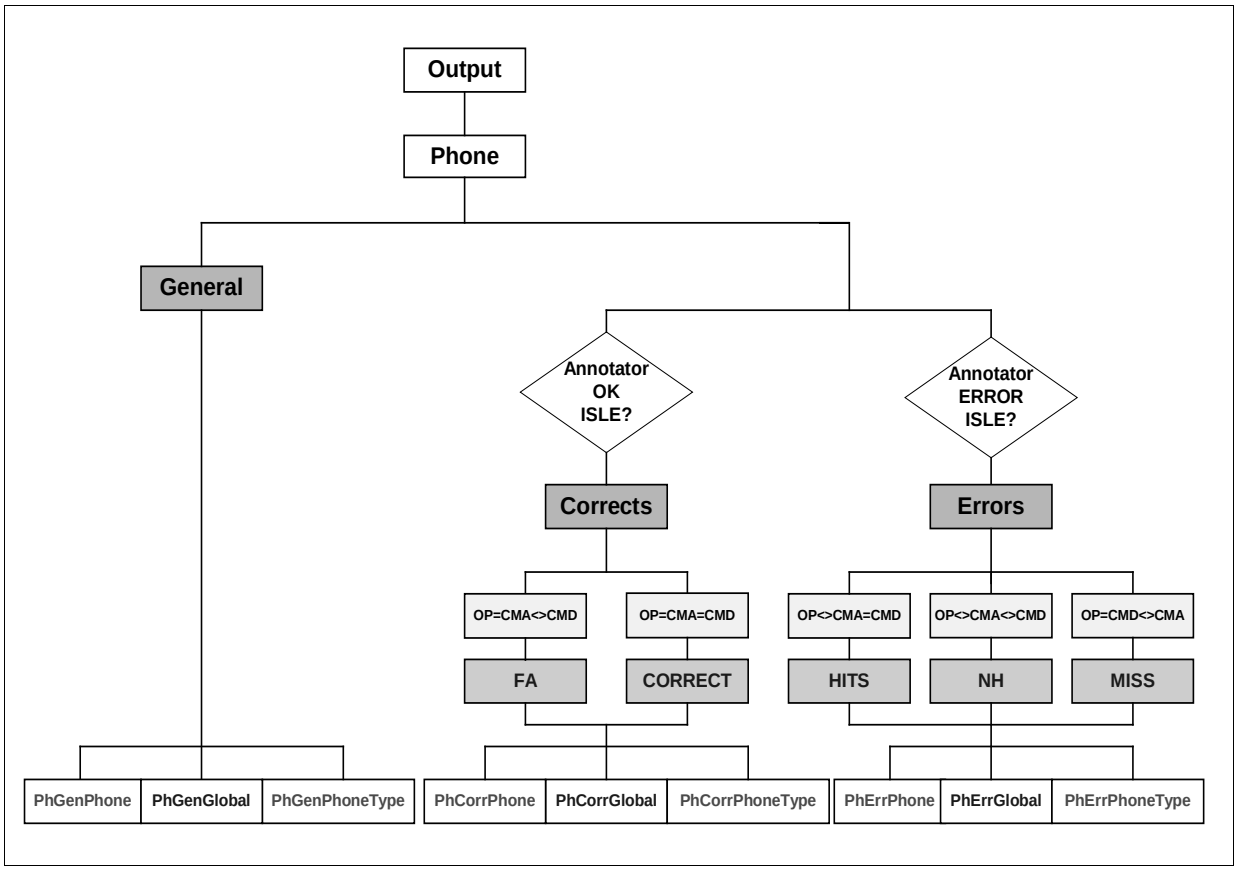


**Figure 2: Output's scheme for phone error analysis**

In the Figure 4 it is described how we built the graphs for the phone analysis.

In particular we consider three groups of graphs:

- **CORRECTS**: this is the situation in which the graphs describe the behavior of ISLE DLLs, when the annotator finds a phone equal to the phone in the .REF file.

  ISLE DLLs can:

  - Recognized the same phone: CORRECT;
  - Recognized a different phone: FALSE ALARM.

- **ERRORS:** it is described the behavior of ISLE DLLs, when the annotator finds an error (a phone different to the phone in the .REF file).

  ISLE DLLs:

  - Find the same error: HITS;
  - Find the different error: NH;
  - Don't find the error: MISS.

- **GENERAL:**

  These graphs are generated with all the values of the HMP variable (see Table 8).

  In Table 11 are reported all the information about the graph.

| LEVEL OF ANALYSIS | TABLE NAME | NAME OF OUTPUT FILE | NAME OF MASTER FILE | HMP'S VALUES USED TO BUILD THE GRAPHIC |
|---|---|---|---|---|
| **GENERAL ANALYSIS** | | | | |
| Global | RptGeneralPhone | PhGenGlobal.xls | PhGeneralMaster.xls | All **HMP**'s values. |
| Phone's Type | RptPhoneType | PhGenPhoneType.xls | PhGeneralMaster.xls | All **HMP**'s values for Phone's Type. |
| Phones | RptPhone | PhGenPhone.xls | PhGeneralMaster.xls | All **HMP**'s values for each phone. |
| **ANALYSIS OF CORRECT PHONES** | | | | |
| Global | RptGenCorrectPH | PhCorrGlobal.xls | PhCorrectMaster.xls | **CORRECT** and **FA**. |
| Phone's Type | RptPHCorrectType | PhCorrPhoneType.xls | PhCorrectMaster.xls | **CORRECT** and **FA** for Phone's Type |
| Phones | RptCorrectPhone | PhCorrPhone.xls | PhCorrectMaster.xls | **CORRECT** and **FA** for each phone. |
| **ANALYSIS OF ERROR PHONES** | | | | |
| Global | RptGenErrorPH | PhErrGlobal.xls | PhErrMaster.xls | **HITS**, **NEAR HITS** and **MISS**. |
| Phone's Type | RptPHErrorType | PhErrPhoneType.xls | PhErrMaster.xls | **HITS**, **NEAR HITS** and **MISS** for Phone's Type. |
| Phones | RptPhoneError | PhErrPhone.xls | PhErrMaster.xls | **HITS**, **NEAR HITS** and **MISS** for each phone. |

**Table 11: The output of the phone error analysis**

#### 4.1.1. Results

| INTERVAL | % |
|---|---|
| CORRECT | 92.7 |
| FA | 7.29 |

| Number of istances | |
|---|---|
| TOTAL | 76690 |
| CORRECT | 71096 |
| FA | 5594 |

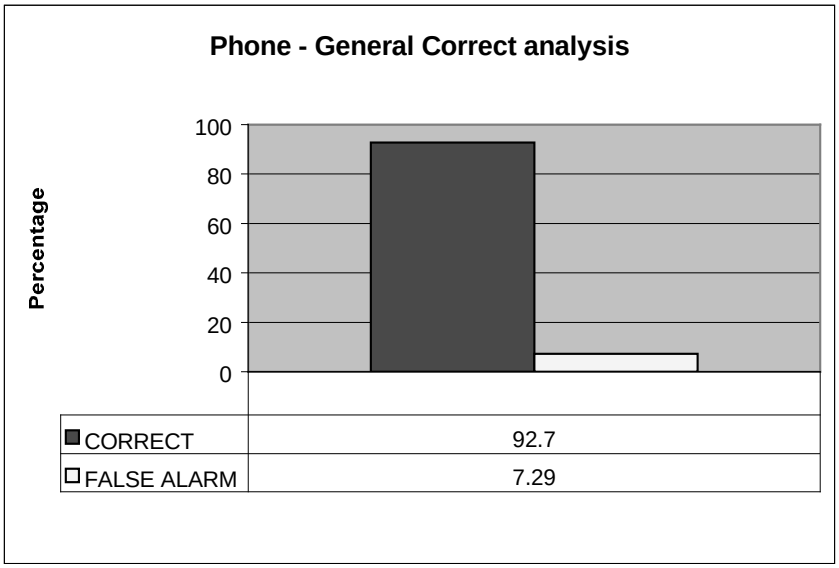


**Figure 3: PhCorrGlobal.xls**

| INTERVAL | % |
|---|---|
| CORRECT | 93.24 |
| FA | 6.75 |

| Number of istances | |
|---|---|
| TOTAL | 711 |
| CORRECT | 663 |
| FA | 48 |

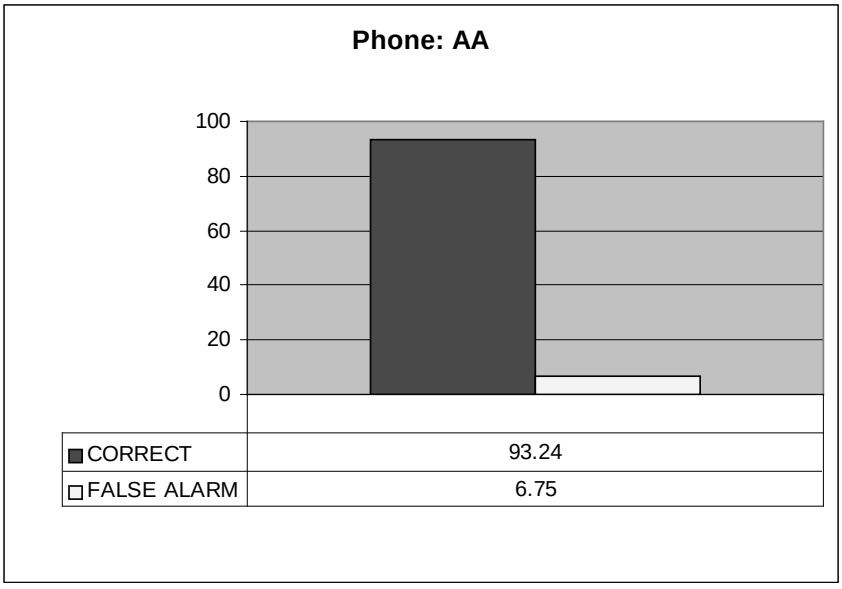


**Figure 4: PhCorrPhone.xls**

| INTERVAL | % |
|---|---|
| CORRECT | 88.77 |
| FA | 11.22 |

| Number of istances | |
|---|---|
| TOTAL | 28628 |
| CORRECT | 25415 |
| FA | 3213 |

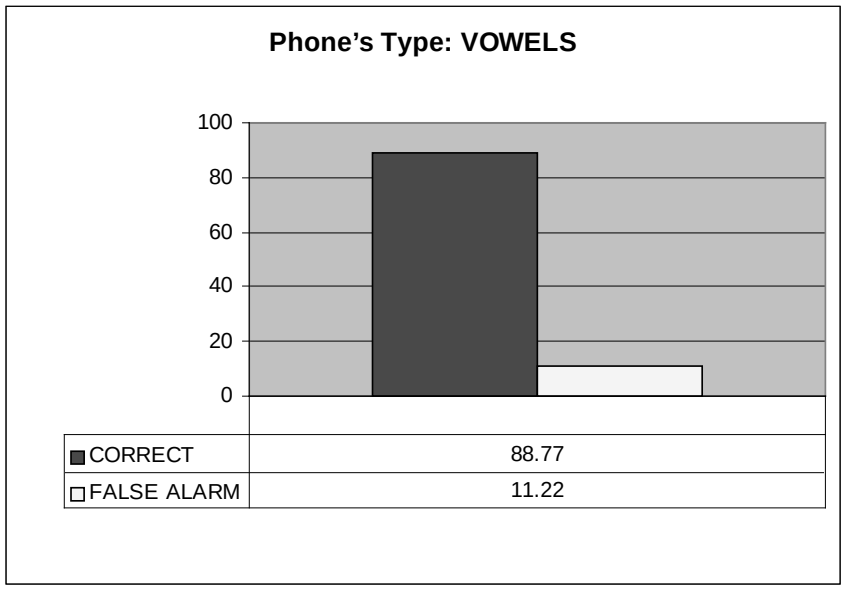


**Figure 5: PhCorrPhoneType.xls**

| INTERVAL | % |
|---|---|
| MISS | 69.08 |
| HITS | 20.06 |
| NEAR HITS | 10.84 |

| Number of istances | |
|---|---|
| TOTAL | 3319 |
| MISS | 2293 |
| HITS | 666 |
| NEAR HITS | 360 |

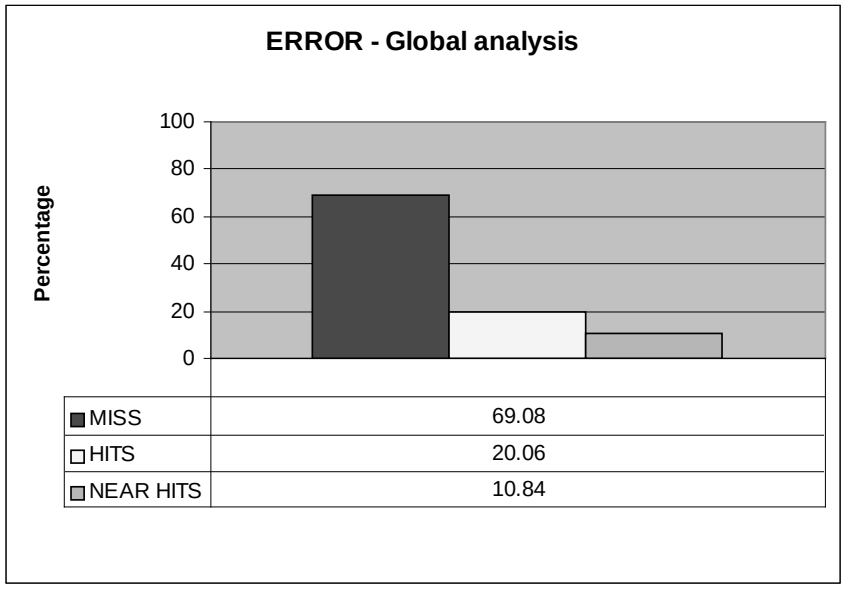


**Figure 6: PhErrGlobal.xls**

| INTERVAL | % |
|---|---|
| MISS | 26.31 |
| HITS | 29.82 |
| NEAR HITS | 43.85 |

| Number of istances | |
|---|---|
| TOTAL | 57 |
| MISS | 15 |
| HITS | 17 |
| NEAR HITS | 25 |

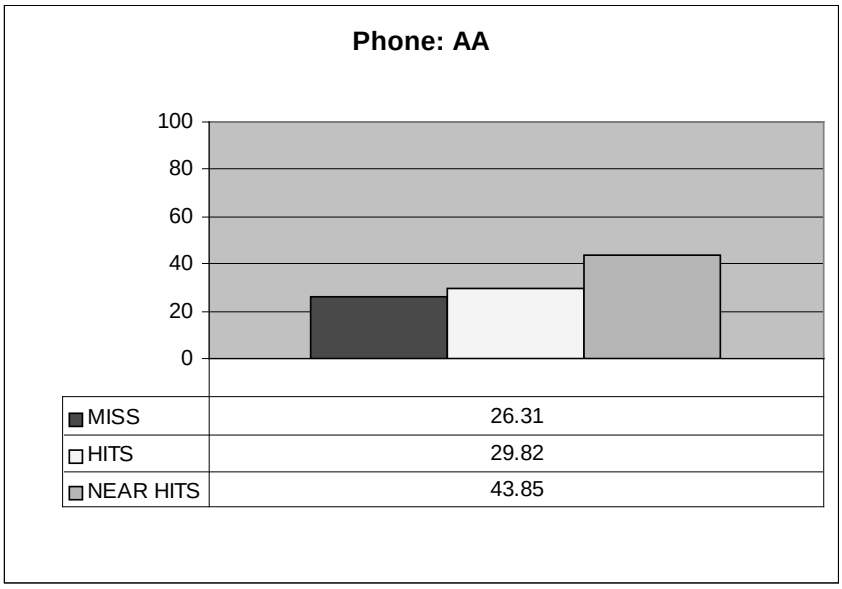


**Figure 7: PhErrPhone.xls**

| INTERVAL | % |
|---|---|
| MISS | 78.67 |
| HITS | 12.05 |
| NEAR HITS | 9.27 |

| Number of instances | |
|---|---|
| TOTAL | 647 |
| MISS | 509 |
| HITS | 78 |
| NEAR HITS | 60 |

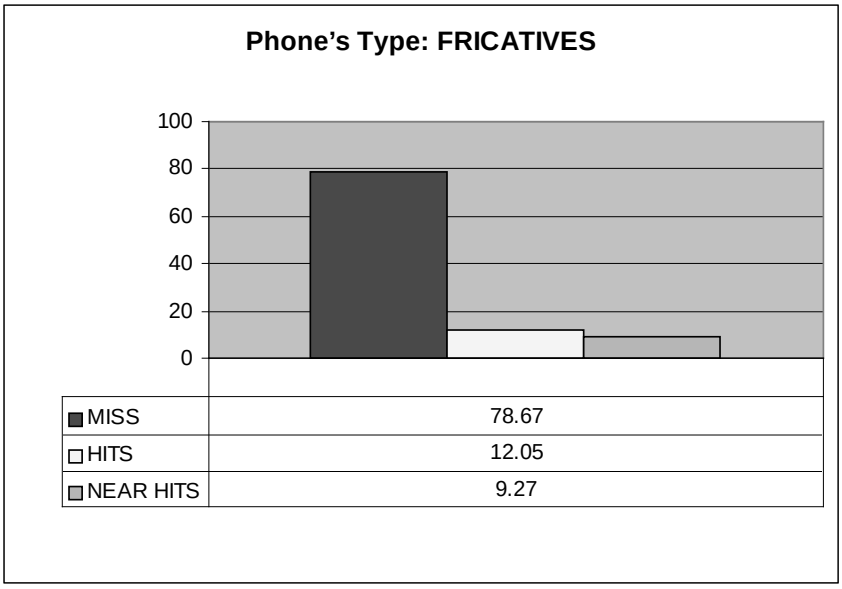


**Figure 8: PhErrPhoneType.xls**

| INTERVAL | % |
|---|---|
| MISS | 2.86 |
| FA | 6.99 |
| HITS | 0.83 |
| CORRECT | 88.86 |
| NEAR HITS | 0.44 |

| Number of instances | |
|---|---|
| TOTAL | 80009 |
| MISS | 2293 |
| FA | 5594 |
| HITS | 666 |
| CORRECT | 71096 |
| NEAR HITS | 360 |

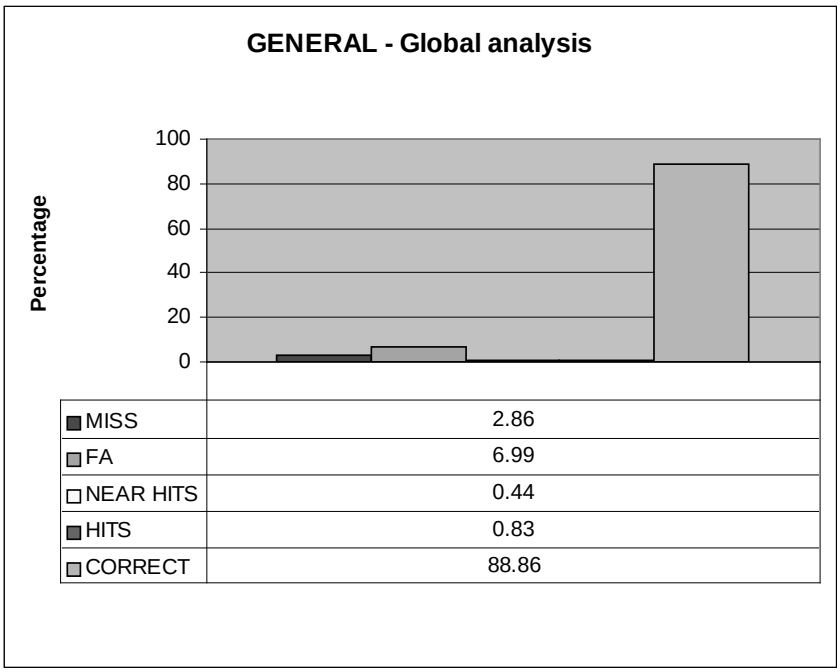


**Figure 9: PhGenGlobal.xls**

| INTERVAL | % |
|---|---|
| MISS | 1.95 |
| FA | 6.25 |
| HITS | 2.21 |
| CORRECT | 86.32 |
| NEAR HITS | 3.25 |

| Number of instances | % |
|---|---|
| TOTAL | 768 |
| MISS | 15 |
| FA | 48 |
| HITS | 17 |
| CORRECT | 663 |
| NEAR HITS | 25 |

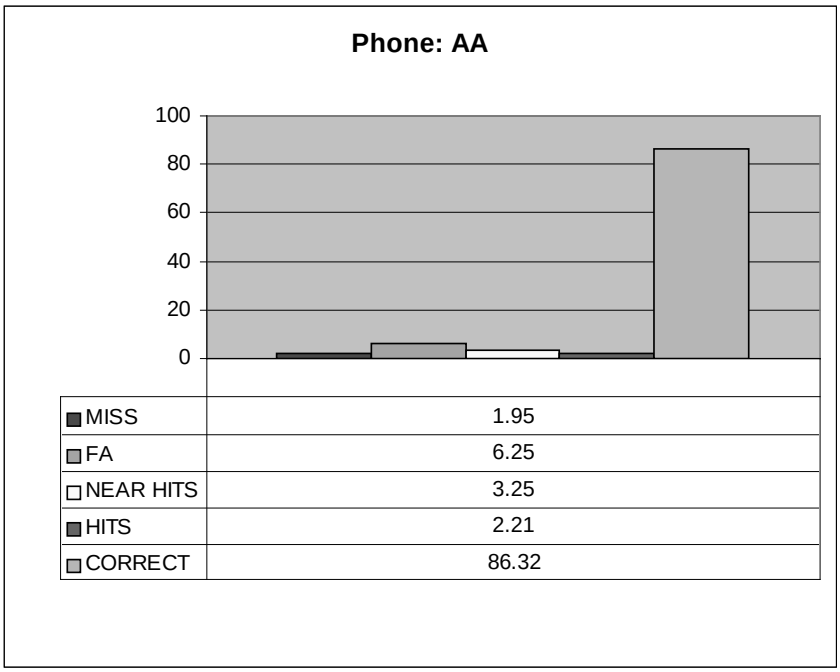


**Figure 10: PhGenPhone.xls**

| INTERVAL | % |
|---|---|
| MISS | 4.88 |
| FA | 3.9 |
| HITS | 0.74 |
| CORRECT | 89.89 |
| NEAR HITS | 0.57 |

| Number of instances | % |
|---|---|
| TOTAL | 10427 |
| MISS | 509 |
| FA | 407 |
| HITS | 78 |
| CORRECT | 9373 |
| NEAR HITS | 60 |

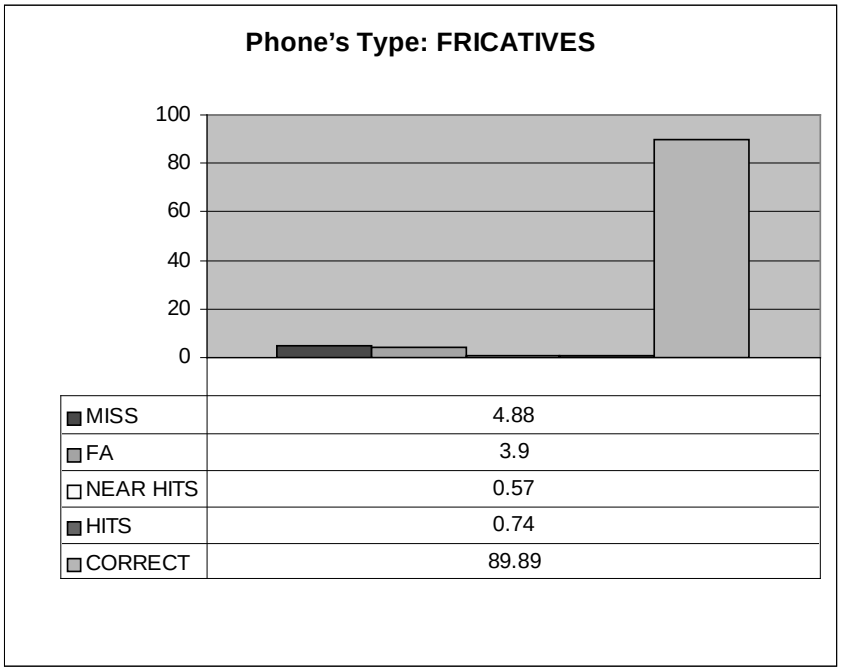


**Figure 11: PhGenPhoneType.xls**

## 4.2. Stress error analysis

### 4.2.1. Phone level

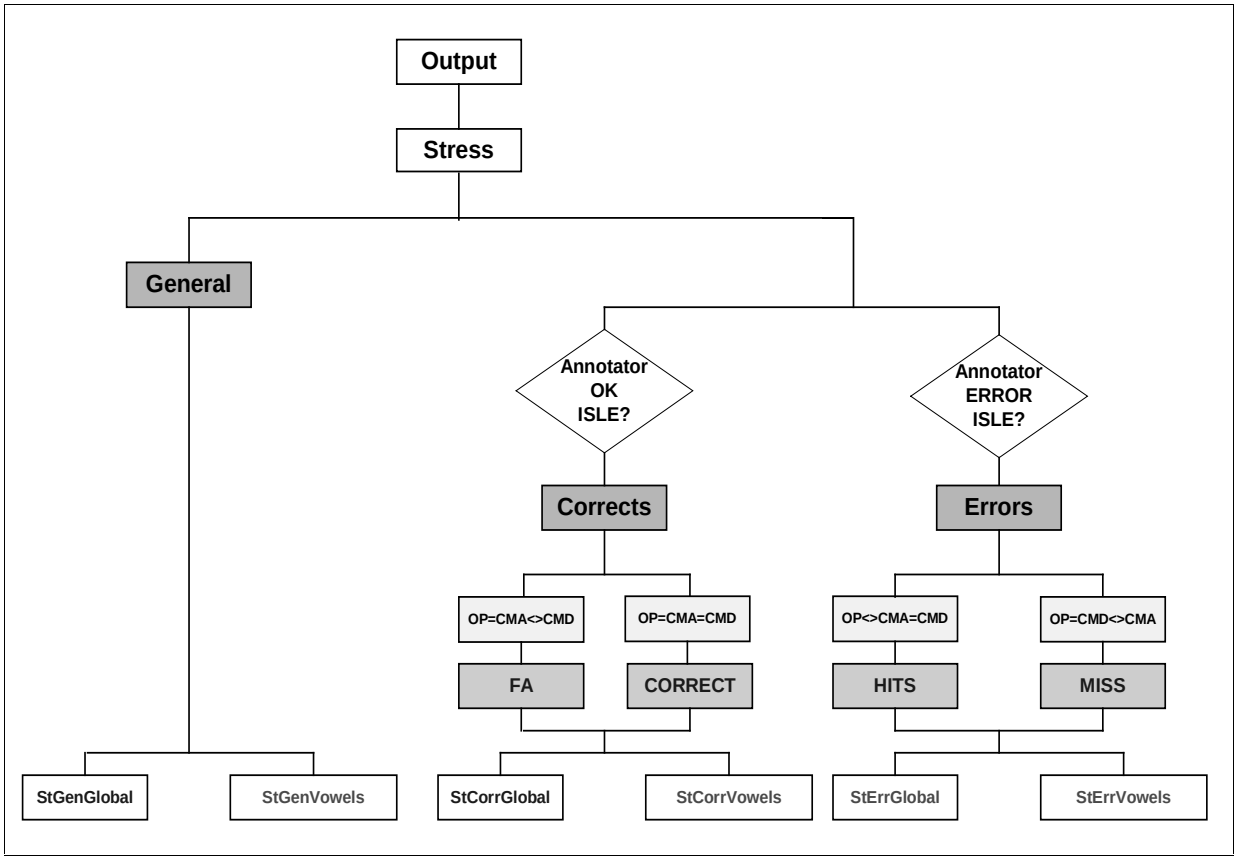


**Figure 12: Output scheme for stress error analysis**

Following the same reasoning used to define phone error analysis (Paragraph 5.1), we generated the graphs for stress errors.
It is important to observe that for the phone-level stress analysis the NH value of the HMS variable is not defined. (See Table 9)

| LEVEL OF ANALYSIS | TABLE NAME | NAME OF OUTPUT FILE | NAME OF MASTER FILE | HMP'S VALUES USED TO BUILD THE GRAPHIC |
|---|---|---|---|---|
| **GENERAL ANALYSIS** | | | | |
| Global | RptGeneralStress | StGenGlobal.xls | StGeneralMaster.xls | All HMS's values. |
| English vowels | RptPhoneStress | StGenVowels.xls | StGeneralMaster.xls | All HMS's values for English vowels. |
| **ANALYSIS OF CORRECTS** | | | | |
| Global | RptGenCorrectStress | StCorrGlobal.xls | StCorrectMaster.xls | **CORRECT** and **FA**. |
| English vowels | RptCorrectStress | StCorrVowels.xls | StCorrectMaster.xls | **CORRECT** and **FA** for English vowels. |
| **ANALYSIS OF ERRORS** | | | | |
| Global | RptGenErrorStress | StErrGlobal.xls | StErrorMaster.xls | **HITS** and **MISS**. |
| English vowels | RptErrorStress | StErrVowels.xls | StErrorMaster.xls | **HITS** and **MISS** for English vowels. |

**Table 12: The output of the stress error analysis**

| PHONE TYPES | |
|---|---|
| **TYPE** | **PHONES** |
| VOWELS | aa ae ah ao aw ax ay eh er ey ih iy oh ow oy uh uw |
| STOP CONSONANTS | p b d t f v g k |
| FRICATIVES | dh th s z sh ch jh zh |
| LIQUIDS | r l m n ng |
| SEMI-VOWELS | y w hh |

**Table 13: Phone types**

#### 4.2.1.1. Results

| INTERVAL | % |
|---|---|
| CORRECT | 98.67 |
| FA | 1.32 |

| Number of instances | |
|---|---|
| TOTAL | 30087 |
| CORRECT | 29689 |
| FA | 398 |

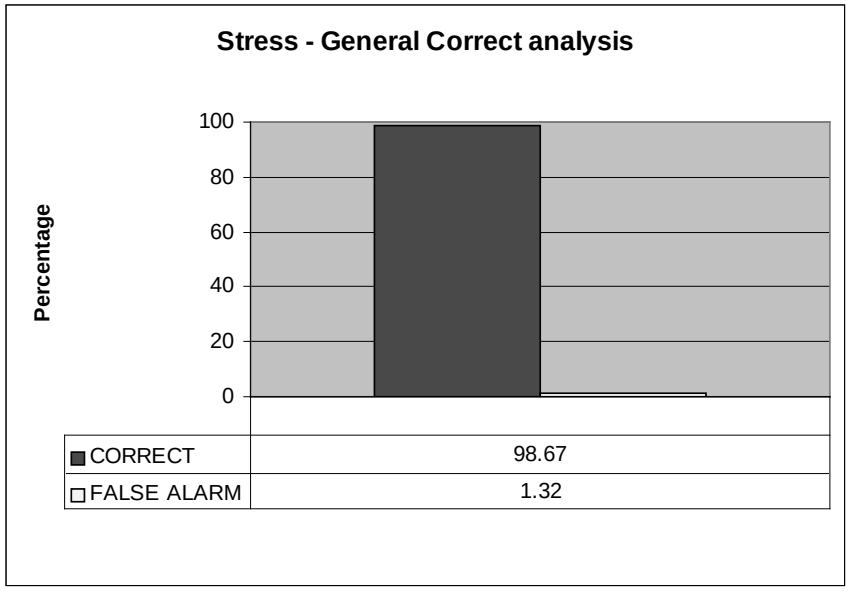


**Figure 13: StCorrGlobal.xls**

| INTERVAL | % |
|---|---|
| CORRECT | 99.21 |
| FA | 0.78 |

| Number of instances | |
|---|---|
| TOTAL | 760 |
| CORRECT | 754 |
| FA | 6 |

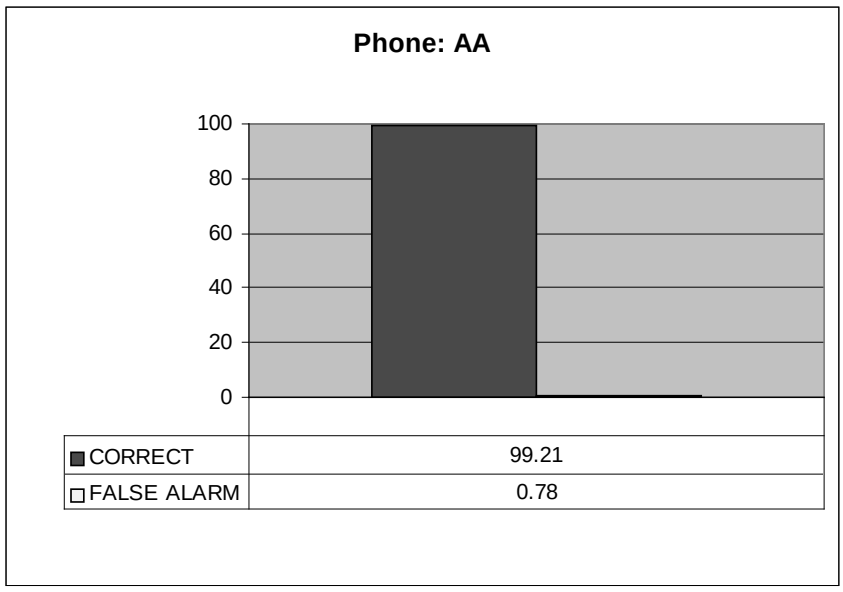


**Figure 14: StCorrVowels.xls**

| INTERVAL | % |
|---|---|
| MISS | 69.51 |
| HITS | 30.48 |

| Number of instances | |
|---|---|
| TOTAL | 410 |
| MISS | 285 |
| HITS | 125 |

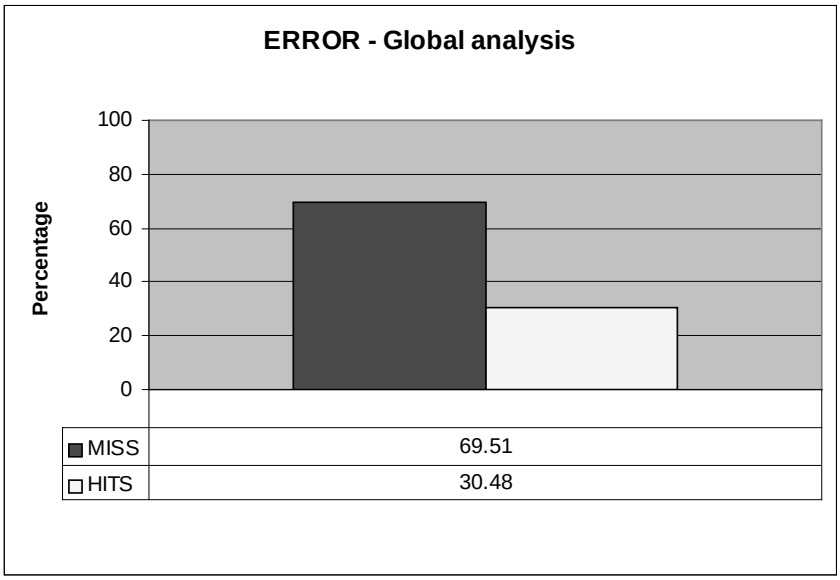


**Figure 15: StErrGlobal.xls**

| INTERVAL | % |
|---|---|
| MISS | 75 |
| HITS | 25 |

| Number of instances | |
|---|---|
| TOTAL | 8 |
| MISS | 6 |
| HITS | 2 |

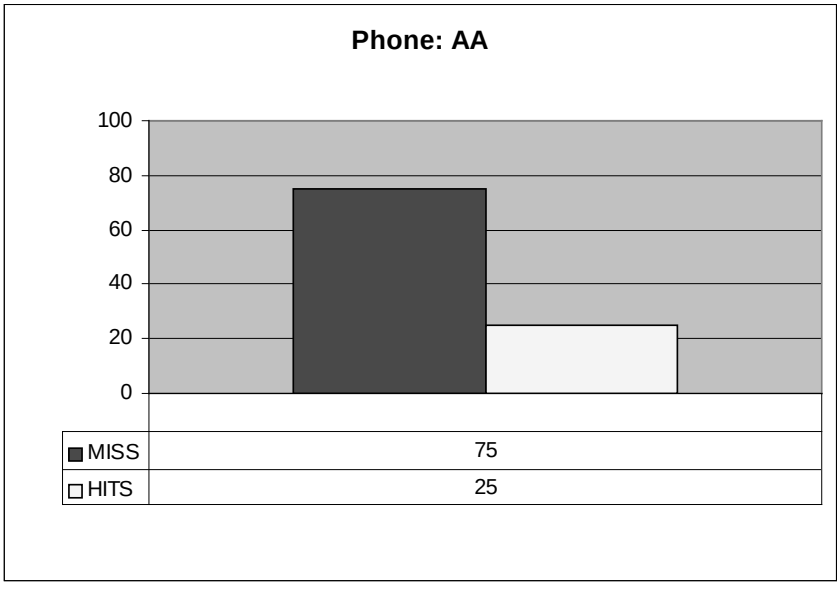


**Figure 16: StErrVowels.xls**

| INTERVAL | % |
|---|---|
| MISS | 0.35 |
| FA | 0.49 |
| HITS | 0.15 |
| CORRECT | 98.98 |

| Number of instances | |
|---|---|
| TOTAL | 79845 |
| MISS | 285 |
| FA | 398 |
| HITS | 125 |
| CORRECT | 79037 |

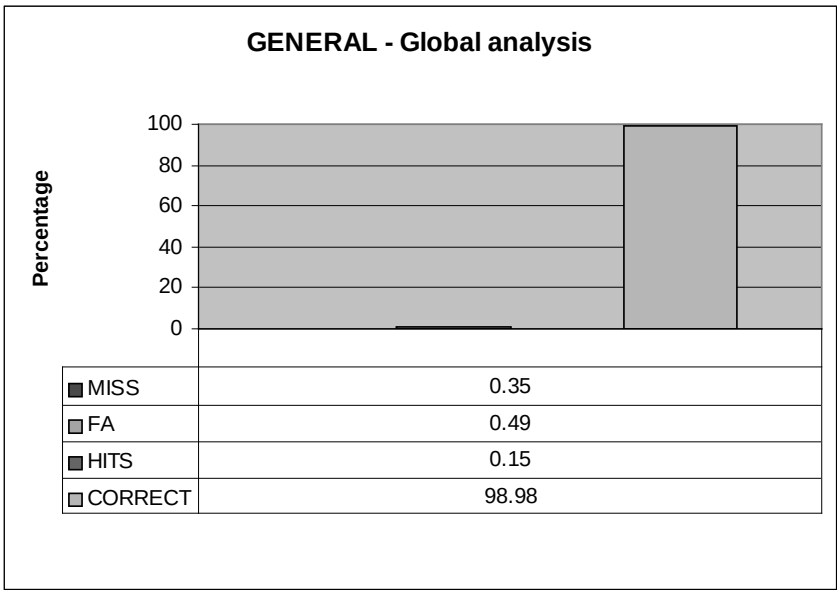


**Figure 17: StGenGlobal.xls**

| INTERVAL | % |
|---|---|
| MISS | 0.78 |
| FA | 0.78 |
| HITS | 0.26 |
| CORRECT | 98.17 |

| Number of instances | |
|---|---|
| TOTAL | 768 |
| MISS | 6 |
| FA | 6 |
| HITS | 2 |
| CORRECT | 754 |

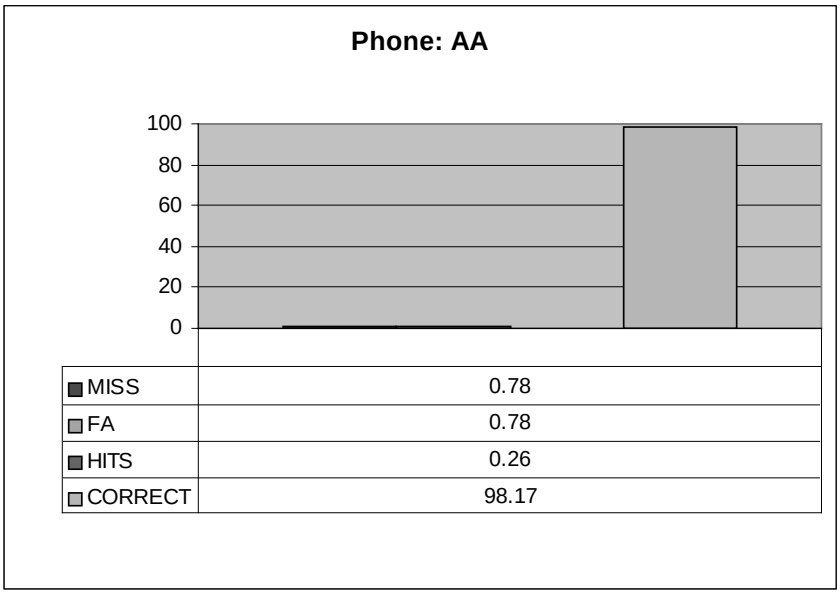


**Figure 18: StGenVowels.xls**

#### 4.2.2. Word level

To generate these results we use an external tool (stress generator tool) that extracts and elaborate the data from the COMPARE table.

### Values of HMS variable for "Word Level"

| KEY | CS | PSA | PSD | HMS |
|---|---|---|---|---|
| **HITS** | X | Y | Y | HITS |
| **NEAR HITS** | X | Y | Z | NH |
| **MISS** | X | Y | X | MISS |
| **FALSE ALARM** | X | X | Y | FA |
| **CORRECT** | X | X | X | |

**Table 14: Values of HMS variable for "Word Level"**

where **X**, **Y**, **Z** are the stress position (value of **OP** variable in Compare Table, see Table 10) in a word.

Example of word level stress analysis:

| Phrase | Wd | CS | PSA | PSD | HMS |
|---|---|---|---|---|---|
| SESS0003_BLOCKE_01 | 1 | 6 | 6 | 6 | |
| SESS0003_BLOCKE_01 | 2 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_01 | 4 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_01 | 5 | 1 | 4 | 1 | MISS |
| SESS0003_BLOCKE_01 | 6 | 1 | 1 | 1 | |
| SESS0003_BLOCKE_02 | 2 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_02 | 3 | 2 | 2 | 4 | FA |
| SESS0003_BLOCKE_02 | 4 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_02 | 8 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_03 | 4 | 3 | 3 | 3 | |
| SESS0003_BLOCKE_03 | 6 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_04 | 2 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_04 | 4 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_04 | 5 | 3 | 3 | 3 | |
| SESS0003_BLOCKE_04 | 6 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_05 | 5 | 7 | 7 | 7 | |
| SESS0003_BLOCKE_05 | 6 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_07 | 2 | 2 | 5 | 2 | MISS |
| SESS0003_BLOCKE_07 | 3 | 6 | 6 | 6 | |
| SESS0003_BLOCKE_07 | 4 | 1 | 1 | 4 | FA |
| SESS0003_BLOCKE_07 | 7 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_08 | 3 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_08 | 5 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_09 | 3 | 2 | 5 | 5 | HITS |
| SESS0003_BLOCKE_11 | 1 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_11 | 2 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_11 | 4 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_11 | 5 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_12 | 2 | 2 | 2 | 2 | |
| SESS0003_BLOCKE_12 | 4 | 1 | 1 | 6 | FA |
| SESS0003_BLOCKE_12 | 6 | 3 | 3 | 3 | |
| SESS0003_BLOCKE_12 | 8 | 1 | 1 | 1 | |
| SESS0003_BLOCKE_13 | 1 | 4 | 6 | 2 | NH |
| SESS0003_BLOCKE_13 | 3 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_13 | 6 | 9 | 9 | 9 | |
| SESS0003_BLOCKE_13 | 7 | 1 | 1 | 1 | |
| SESS0003_BLOCKE_13 | 9 | 4 | 4 | 4 | |
| SESS0003_BLOCKE_14 | 1 | 1 | 1 | 3 | FA |
| SESS0003_BLOCKE_14 | 8 | 5 | 5 | 5 | |
| SESS0003_BLOCKE_15 | 7 | 3 | 3 | 3 | |
| SESS0003_BLOCKE_34 | 6 | 4 | 4 | 1 | FA |

**Table 15 : An example stress table on the word level**

In this pictures the Rate are calculated in this way (# means number of):

| FA RATE | HITS RATE | NH RATE |
|---|---|---|
| **#FA / ( #FA + #CR )** | **#HITS / ( #HITS + #NH + #MISS )** | **( #HITS + #NH ) / ( #HITS + #NH + #MISS )** |

**Table 16: Rate formulas**

#### 4.2.2.1. Results

| HMS | Total | CountOfHMS |
|---|---|---|
| CR | 6040 | 5233 |
| FA | 6040 | 398 |
| HITS | 6040 | 125 |
| MISS | 6040 | 282 |
| NH | 6040 | 2 |

| FA RATE | HITS RATE | NH RATE |
|---|---|---|
| 7.07% | 30.56% | 31.05% |

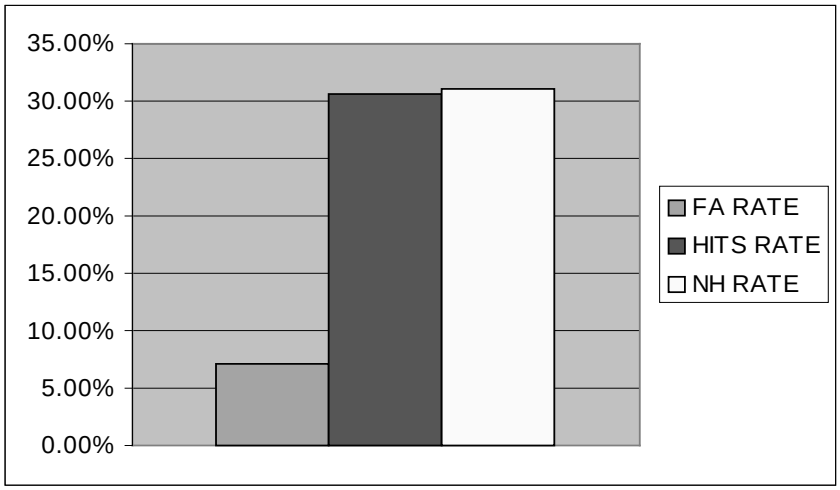


**Figure 19: Results "Word Level" Stress for German speakers**

| HMS | Total | CountOfHMS |
|---|---|---|
| CR | 5210 | 4277 |
| FA | 5210 | 531 |
| HITS | 5210 | 100 |
| MISS | 5210 | 297 |
| NH | 5210 | 5 |

| FA RATE | HITS RATE | NH RATE |
|---|---|---|
| 11.04% | 24.88% | 26.12% |

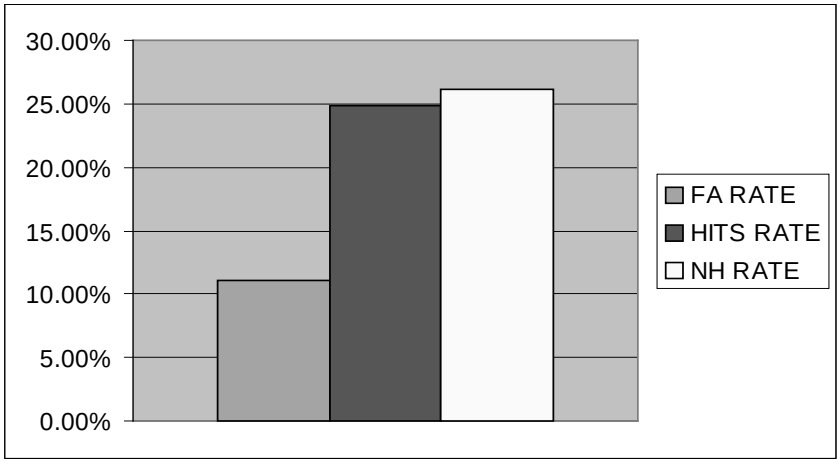


**Figure 20: Results "Word Level" Stress for Italian speakers**

# 5. Experiments with the recognition threshold

Recognition with IHAPI can have one of two results:
1. It recognizes and so aligns the speech signal with the some path through the syntax.
2. It fails to recognize.

The input can also be:
1. GOOD: its matches what IHAPI has been told to recognize;
2. BAD: the words in the utterance are different or this is nonsense noise.

So there are the following possibilities:

| | Utterance matches prompt | Utterance is bad or pure noise |
|---|---|---|
| **IHAPI aligns utterance to prompt** | HIT | FALSE ACCEPT |
| **IHAPI fails to align utterance to prompt** | MISS | CORRECT REJECT |

**Table 17: IHAPI Alignment**

In order to get as many HITS as possible, the ISLE demonstrator uses an adaptation process, so that the recognizer is better able to handle the differences between the student's speech and the trained, UK models. The problem is to avoid FALSE ALARMS and MISSES: clearly, if the recognizer is made very strict, it will reject almost every utterance, giving as results very few FALSE ACCEPTS, but also very many MISSES.
To do this it computes the "average word confidence" across the sentence after recognition and if this value is below some threshold, we pretend (to the top level) that recognition failed.
Thus even if the recognizer successfully aligns

*'fuffa fuffa'*

with the prompt

*"they asked if I wanted to come along on the barge trip",*

we should be able to reject the utterance.

For each sentence we will then get two average confidence values:

- Correct-prompts
- Wrong-prompts

These confidence values are available after the recognition stage:

**if recognition fails, output a zero for that trial, otherwise the utterance is used.**

We generate four types of ‘incorrect’ prompts:

1. **Non-sense phrase as syntax**
   Example:
   **Syntax**: ANYTHING COFFEE MANY WAY TRADITIONAL
   **Wav file**: I SAID THROUGH NOT THOUGH

2. **Two words exchanged one another**
   Example:
   **Syntax**: HE HAS HIS OWN STUDIO PHOTOGRAPHIC
   **Wav file**: HE HAS HIS OWN PHOTOGRAPHIC STUDIO

3. **1 word repeated twice**
   Example:
   **Syntax**: A STUDENT VISA PERMITS PERMITS THEM TO STAY LONGER
   **Wav file**: A STUDENT VISA PERMITS THEM TO STAY LONGER

4. **BAD SESSIONS:**
   These ‘incorrect’ prompts are realistic errors that people make when reading sentences.
   Our Italian/German speakers were asked to record hundreds of prompts. In most cases, they read the prompt word-for-word as expected.
   Sometimes, though, they inserted, deleted, or repeated words, or in other ways mangled the sentences.
   Thus, these ‘incorrect’ prompts are subsets of our corpus, for which the original (expected) and the actual (corrected) prompts are different.

   Example 1:
   **Syntax**: SAID THROUGH NOT THOUGH
   **Wav file**: I SAID THROUGH NOT THOUGH

   Example 2:
   **Syntax**: SINGERS LEARN HOW TO PROJECT THEIR VOICES
   **Wav file**: SINGERS LEARN HOW TO PROTECT THEIR VOICES

## 5.1 Results

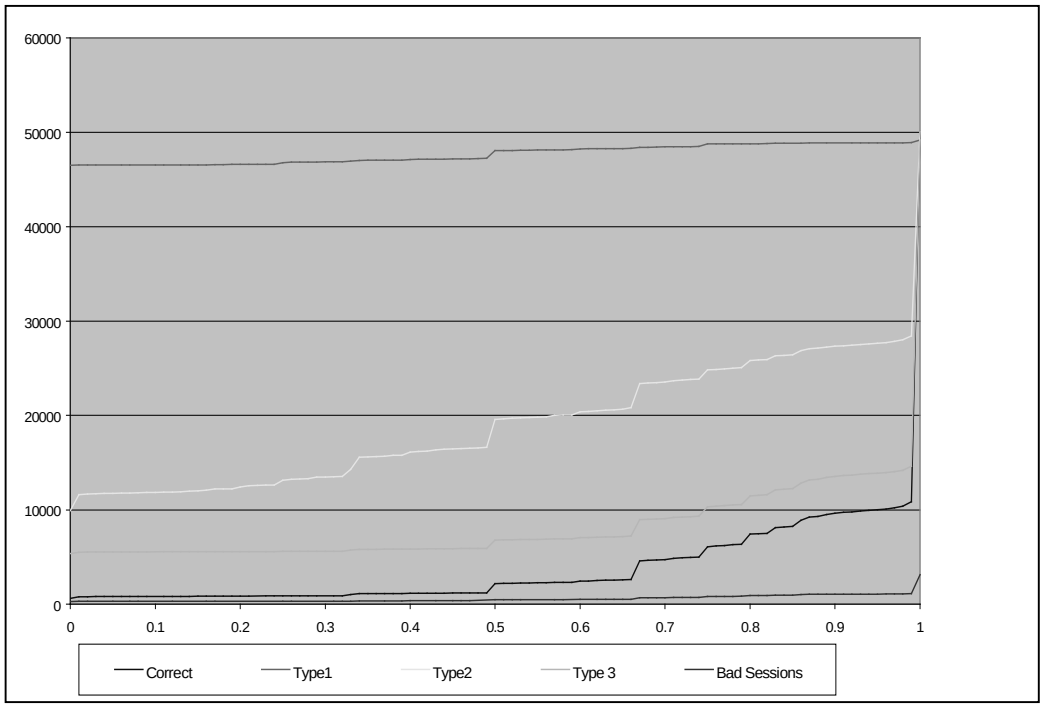


**Figure 21 : Cumulative graph**

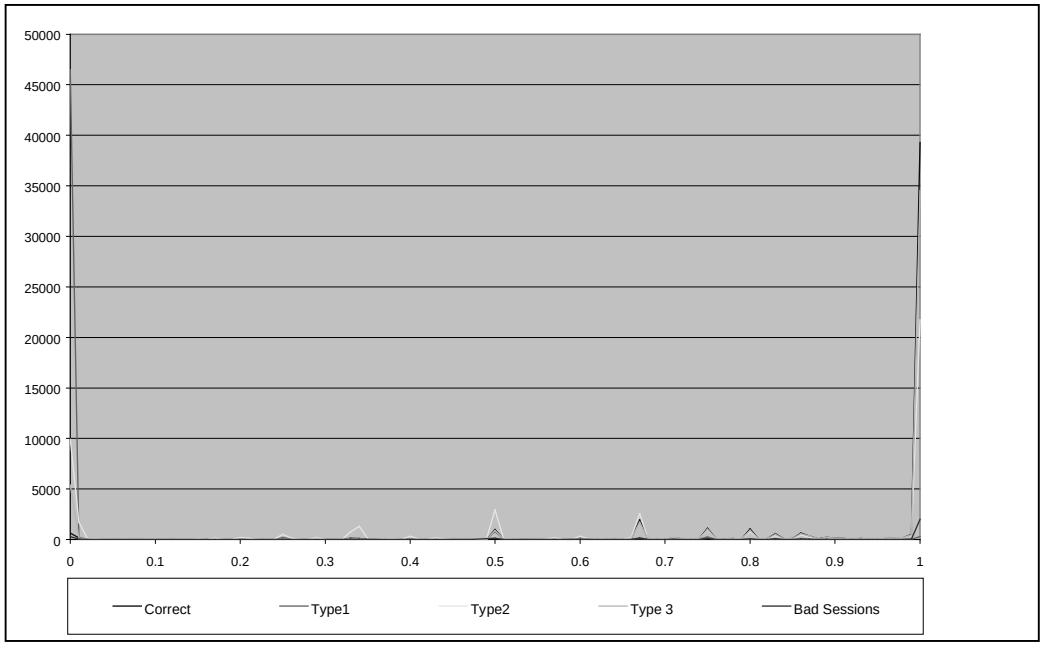


**Figure 22: Frequencies graph**

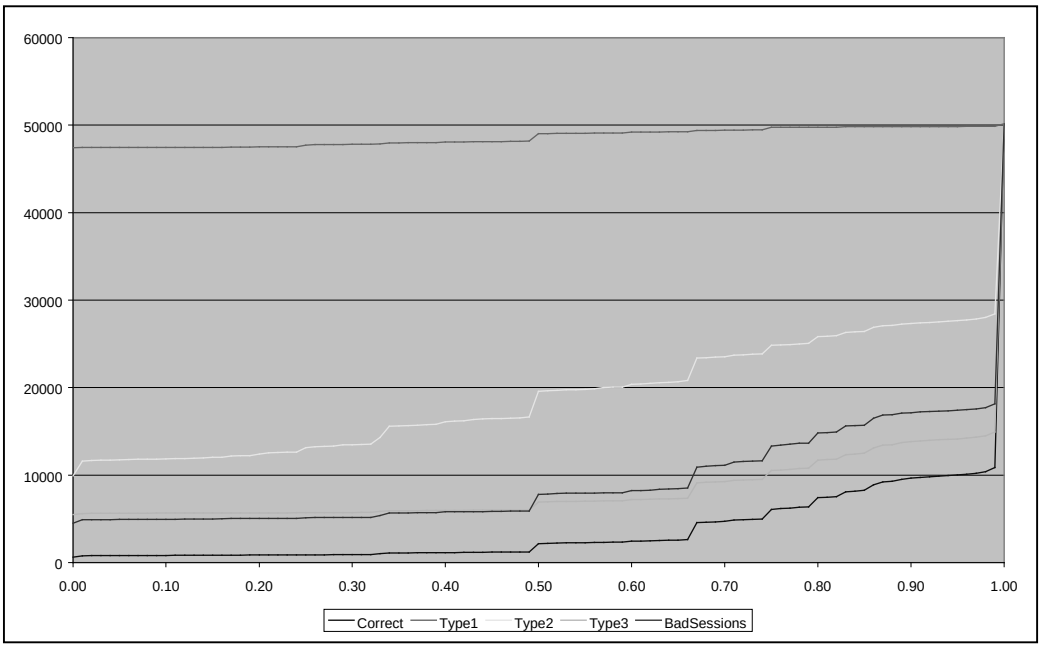


**Figure 23: Cumulative percentage graph**

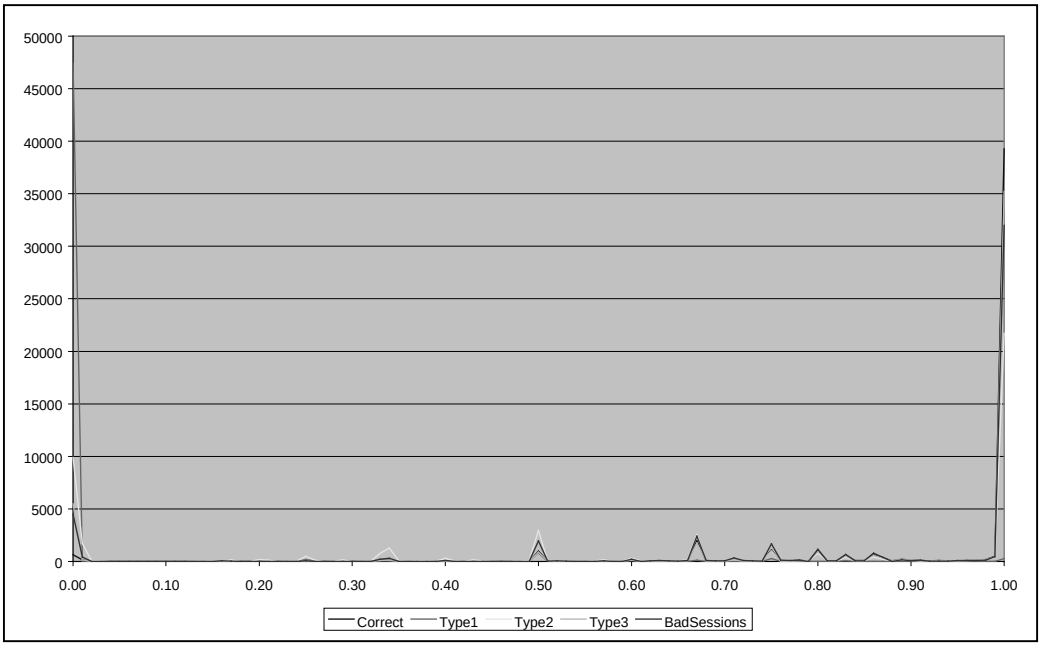


**Figure 24: Frequencies percentage graph**

# 6. Experiments with the localization threshold

The question we want to ask is how well the system is able to find words/phones with errors. Errors are defined as words/phones that the annotators scored as incorrect.
Because in the current demonstrator we only highlight entire words and not single phones, we limit these tests to trying to find a threshold that lets us automatically find as many of the words with real errors as possible, while mis-localizing as few as possible of the 'good' words.

In the validation tool the localization process is carried out in a second pass, after the recognition. In the recognition stage (unless we recognize the utterance), we will determine the sequence of words spoken by the student. We will then re-recognize the same audio file (.wav file), allowing *only that sequence of words* as in the prompt (i.e., in a multiple choice exercise, the recognition decides between various answers, but localization only focuses on the one spoken by the student.)

Localization will also use adapted models, but not the phone-level adapted models used in recognition. In recognition we don't care *how* the student spoke, only *what* she spoke.

In localization we want to know how well she spoke the words, and thus we do *not* want to use models that 'make it easier' on the student by allowing for differences between her pronunciation and the target UK accent. Nevertheless we would like to eliminate the variability due to microphone, room conditions, and general properties of her vocal apparatus.

From the validation point of view all we want to know is:

**What threshold best distinguishes between those words that were somehow wrong according to the human annotation and those that were ok?**

So, we should recognize a large number of sentences, using only the correct prompt this time, and extract the localization confidence scores for words and phones. In the case of word confidences, we generate a table like the one in Table 18.
For phones, you could have a similar table, but the "OK"/"BAD" decision could refer either to the phone or to the entire word (the latter is simpler to analyze and more immediately useful to us, but less interesting.) The aim of all this is to find the best threshold, i.e., the one that is greater than most "BAD" words and smaller than most "OK" words.

| Word confidence | Annotator |
|---|---|
| 0.78732244 | OK |
| 0.721430742 | OK |
| 0.901270757 | OK |
| 0.742801423 | BAD |
| 0.913641704 | OK |
| 0.583372004 | BAD |
| 0.672914373 | BAD |
| 0.571764351 | BAD |
| 0.745568447 | OK |
| 0.795128121 | OK |
| 0.804217554 | OK |
| 0.751592599 | BAD |

**Table 18: Word confidence annotation**

## 6.1 Results

Comparing the confidence scores for “OK” and “BAD” words (see Table 18) we generate the following figures for the word and the phone level.

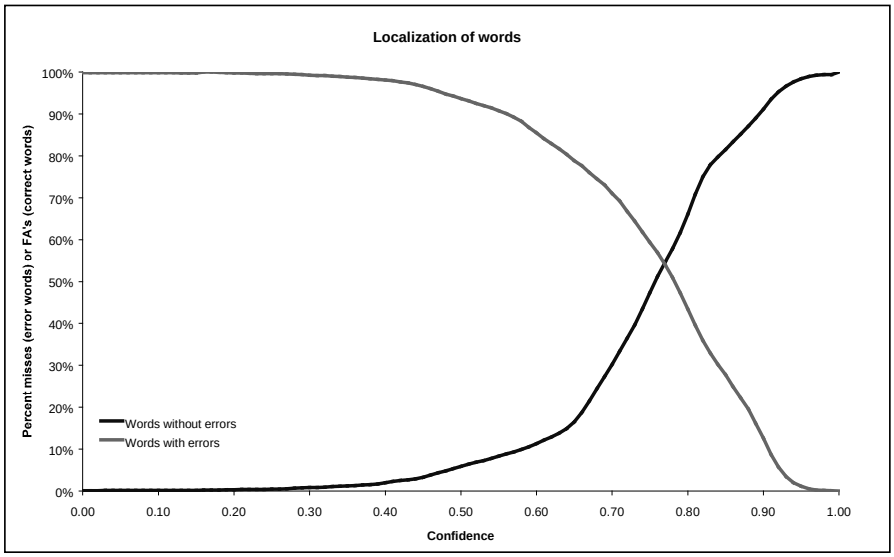


**Figure 25: Word level localization threshold**

The phone-level threshold could even be specific to particular phones or classes of phones, although this should not, in theory, make much difference using the new Gaussian classifiers.

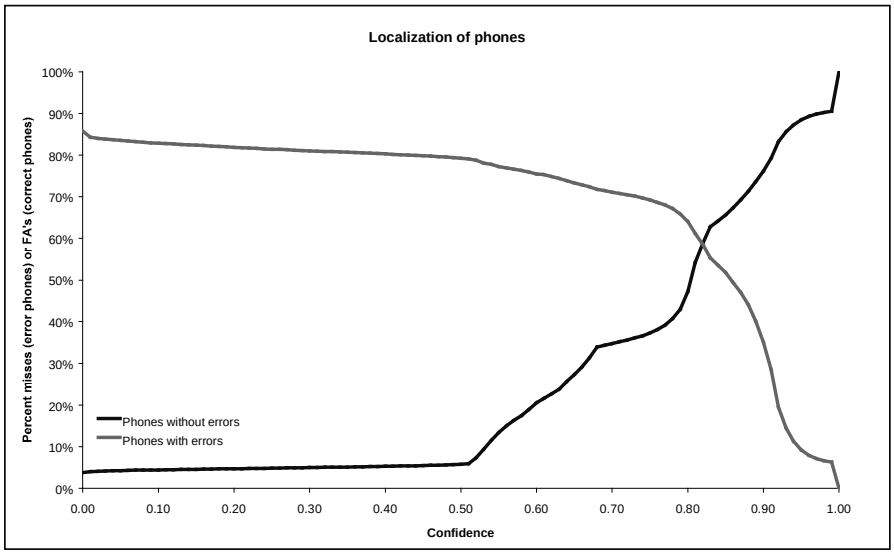


**Figure 26: Phone Level localization threshold**

# Appendix 1: On-line Evaluation: Instructions for the evaluator

1. try to ensure low background noise and distractions
2. allow the user to open the program from closed
3. observe and note down any problems experienced by the user
4. note down any reactions of the user (visual or verbal)
5. allow users to proceed through the program as they wish
6. allow users to stop and exit when they want
7. note down the components of the program used
8. record the total time spent
9. complete the questionnaire with the user after the session

# Appendix 2: Introductory information

ISLE is a 2-year project funded by the EU (ending in March 2000), which aims to develop computer-based training for language learners wishing to improve their pronunciation. The main features are:

- it is designed for Italian and German adult learners of English,
- it is designed for self-study,
- it uses speech recognition technology to identify pronunciation errors,
- it identifies where the error has occurred in an utterance and what kind of error it is,
- it provides verbal feedback on the screen, not just an accuracy score (eg "You said 'beet' instead of 'bit'"),
- it offers further practice of specific sounds related to the error (eg minimal pairs),
- it provides a variety of exercises for pronunciation practice (dialogues, listen and repeat, read and repeat, question and answer, sentence building),
- it has a multi-media graphical design.

The version that is being evaluated at the end of the project is for demonstration purposes only. It is not intended for sale and is far from being a marketable piece of software. The recording of the dialogues, for example, was not made under professional conditions. The design was mainly determined by technical not pedagogical considerations. The aim of the on-line evaluation procedure is to get the reaction of learners and teachers towards the above features, rather than its value as a finished product.

For each evaluation session an evaluator will be present to

- ensure that the program runs smoothly,
- observe what the user is doing and fill in an evaluator's sheet,
- administer the user's questionnaire and elicit feedback on the session.

Each session should take a minimum of ¾ hour to use the program and ¼ hour for the questionnaire.

## Appendix 3: Evaluator's record sheet

Location: _____

Name of user: _____ Name of evaluator: _____

Date of session: _____ Time started: _____ Time ended: _____

Problems experienced by the user:

_____

_____

_____

Reactions of the user (visual or verbal):

_____

_____

_____

Components of the program used:

_____

_____

_____

_____

Other comments:

_____

_____

_____

# Appendix 4: Sessions analyzed during the off-line evaluation

| German Sessions | | | Italian Sessions | | |
|---|---|---|---|---|---|
| **SESSION NAME** | **SPEAKER SEX** | **PARTNER** | **SESSION NAME** | **SPEAKER SEX** | **PARTNER** |
| SESS0006 | Female | ULeeds | SESS0003 | Male | Dida El. |
| SESS0011 | Male | ULeeds | SESS0040 | Male | Dida El. |
| SESS0012 | Male | ULeeds | SESS0041 | Female | Dida El. |
| SESS0015 | Male | ULeeds | SESS0121 | Male | UMilan |
| SESS0020 | Male | ULeeds | SESS0122 | Female | UMilan |
| SESS0021 | Female | ULeeds | SESS0123 | Male | UMilan |
| SESS0161 | Male | UHam | SESS0124 | Male | UMilan |
| SESS0162 | Male | UHam | SESS0125 | Male | UMilan |
| SESS0163 | Female | UHam | SESS0126 | Male | UMilan |
| SESS0164 | Male | UHam | SESS0127 | Male | UMilan |
| SESS0181 | Female | Klett | SESS0128 | Female | UMilan |
| SESS0182 | Male | Klett | SESS0129 | Female | UMilan |
| SESS0183 | Female | Klett | SESS0131 | Male | UMilan |
| SESS0184 | Female | Klett | SESS0130 | Male | UMilan |
| SESS0185 | Male | Klett | SESS0132 | Male | UMilan |
| SESS0186 | Male | Klett | SESS0133 | Male | UMilan |
| SESS0187 | Male | Klett | SESS0134 | Male | UMilan |
| SESS0188 | Male | Klett | SESS0135 | Male | UMilan |
| SESS0189 | Male | Klett | SESS0136 | Male | UMilan |
| SESS0190 | Female | Klett | SESS0137 | Male | UMilan |
| SESS0191 | Female | Klett | SESS0138 | Male | UMilan |
| SESS0192 | Female | Klett | SESS0139 | Male | UMilan |
| SESS0193 | Male | Klett | SESS0140 | Male | UMilan |

**Table 19 : The sessions**