5 CoDIT: A combined discrimination/identification tool for speech perception experiments

Hartmut R. Pfitzinger

5.1 Introduction

Scientific tools for speech research are rather rarely presented even though well-known exceptions exist e.g. *Praat* (Boersma & van Heuven 2001). Regarding speech perception test tools only one was found (Thon 1982). In the past, perception tests have predominantly been conducted with printed answer sheets while the benefits of computer-aided assessment tools are pretty obvious. This paper presents the tool *CoDIT (Combined Discrimination/Identification Tool)*, shows past applications, and discusses advantages and disadvantages as well as when to use it and when not to.

5.2 Origins

In my first perception experiments (Pfitzinger 1994) 20 participants drew crosses on a total of 80 printed A4 answer sheets, each with 60 small cardinal vowel diagrams, in order to be able to determine the respective perceived vowel quality as accurately as possible. Then I had to read the judgments with a mask (Fig. 5.1) and enter the 9600 measured values into the computer. Inaccuracy and susceptibility to errors could only be reduced with considerable effort. Even before the experiments, it was clear to me that a new interactive perception test tool was necessary: *"This method should be used in future perception tests as it offers clear advantages:*

- *Individual test speed*
- *Output of perception results in machine-readable form*
- *Availability of additional information (How often was a stimulus played? How long did the respective judgment take? How often was it refined? . . .)*
- *Eliminate errors caused by manual reading of judgments*
- *Maximum precision and accuracy*

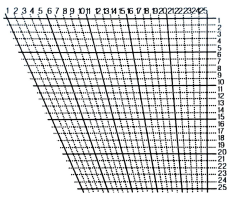

Fig. 5.1: Original-size mask for reading of the two dimensions of a marker in a printed cardinal vowel diagram on an answer sheet.

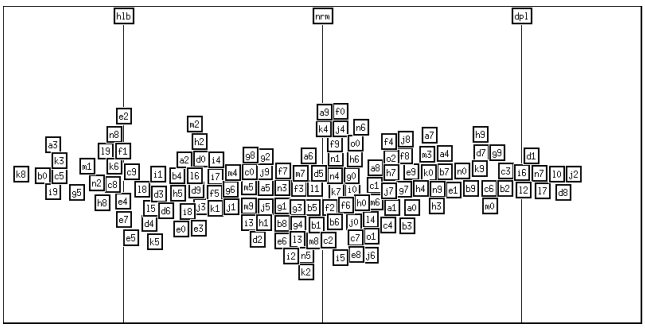

Fig. 5.2: "The completed answer sheet of one subject in the [local speech rate] perception test." (Pfitzinger 1998)

- *Possibility of a preceding individual training or test phase based on results that have already been validated" (Pfitzinger 1994, 48)*

But only a perception experiment to assess the speech rate of short stimuli (Pfitzinger 1998), in which a total of 60 subjects took part, required exactly this perception test tool, which I developed in 1998 using the X Window System and OSF/Motif. Fig. 5.2 shows the result of one of the first five participants from 1998. A short time later, 29 subjects had already carried out the test, so that I could gain a lot of experience and improve the tool even further (Pfitzinger 1999).

Thus, between 1999 and 2003, 40 subjects could participate in a new computer-aided version of the phonetic vowel quality perception test. Fig. 5.3 and 5.4 show the graphical user interface before and after an example participant un-

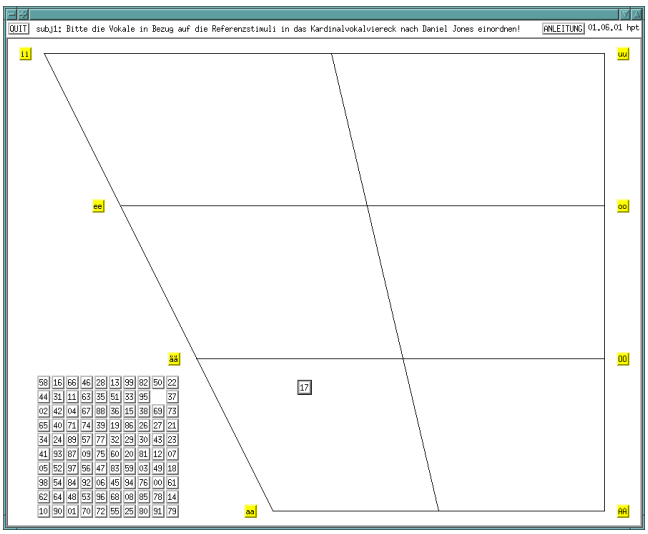

Fig. 5.3: Here the user interface of CoDIT was configured for conducting phonetic vowel quality perception tests (Pfitzinger 2003).

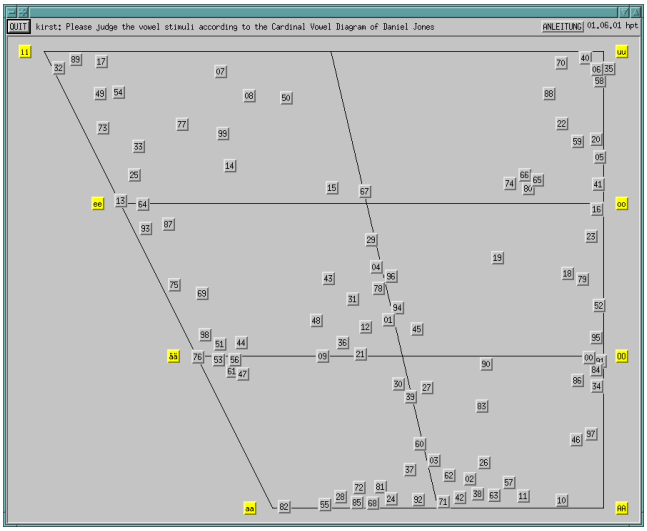

Fig. 5.4: The stimuli were shifted from its initial positions into the response area (cf. Fig. 5.3).

dertook the experiment. Tillmann & Pfitzinger (2004) first mentioned the tool as a useful speech perception research tool, albeit with a temporary name.

The last major upgrade was necessary in 2009 for the perception experiment of Willing (2010) and consisted of an option to additionally include *.bmp*-images at desired positions in the user interface (Fig. 5.5).

5.3 Designing an experiment

A CoDIT experiment is defined by the test stimuli and optional reference stimuli and pictures. A configuration file contains all information CoDIT needs to draw the graphical user interface and to load the test stimuli beforehand.

Table 5.1 shows an example configuration file that defines an experiment similar to Fig. 5.5, and Table 5.2 lists all possible commands of a configuration file and explains them.

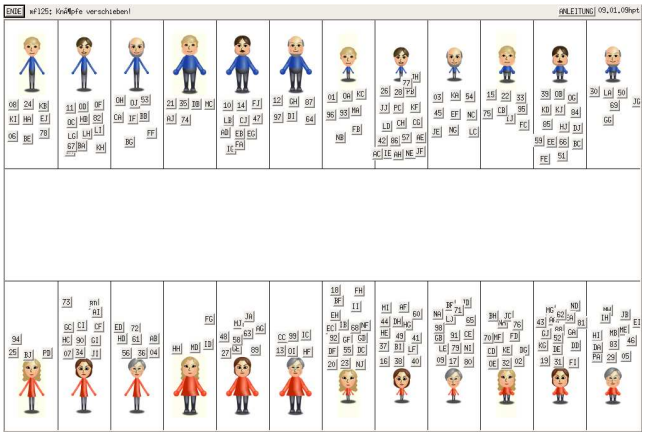

Fig. 5.5: In this experiment speech stimuli had to be assigned to avatars representing male/female, young/middle-aged/old, tall/small, and overweight/underweight persons (Willing 2010).

The test stimuli are not defined in the configuration file but in an extra stimulus definition and response file (Tab. 5.3) for three reasons:

1. In perception experiments the test stimuli quite often need to be revised before achieving the finally convincing version. Until then, editing only this file is necessary.
2. This file stores the actual positions of all test stimuli when leaving the experiment e.g. for a break. Thus, when restarting CoDIT, the stimuli are at the most recent positions.
3. This file is easy to read (e.g. with *Matlab, R, Excel*, ...) as it contains the test stimulus labels, filenames and its x/y-coordinates.

```
; Experiment: assign speech stimuli to pics of speakers
N G
; Subjects using this account have to type-in a unique
; 5-letter ID which CoDIT appends to the .dat-filename:
U 10000
H Knöpfe verschieben!
O 10
V
S 48000.
M 150000
D /home/subjects/exp_spk_pic/instructions.doc
F
T /home/subjects/exp_spk_pic/dat/stimuli.dat
G /home/subjects/exp_spk_pic/log/spk_pic.log
X 1000
Y 650
;; Two horizontal lines separate upper and lower parts
L 0 275 1000 275
L 0 375 1000 375
;; Vertical lines separate the upper part in target areas
L 125 0 125 275
L 250 0 250 275
L 375 0 375 275
L 500 0 500 275
L 625 0 625 275
L 750 0 750 275
L 875 0 875 275
L 1000 0 1000 275
;; Vertical lines separate the lower part in target areas
L 125 375 125 650
L 250 375 250 650
L 375 375 375 650
L 500 375 500 650
L 625 375 625 650
L 750 375 750 650
L 875 375 875 650
L 1000 375 1000 650
;; 8 upper pictures of 8 male speakers
W  62 70 ./pic/mglA1_S.bmp
W 187 70 ./pic/mglA1_S.bmp
W 312 70 ./pic/mglA1_S.bmp
W 437 70 ./pic/mglA1_S.bmp
W 562 70 ./pic/mglA1_S.bmp
W 687 70 ./pic/mglA1_S.bmp
W 812 70 ./pic/mglA1_S.bmp
W 937 70 ./pic/mglA1_S.bmp
;; 8 lower pictures of 8 female speakers
W  62 580 ./pic/mglA1_S.bmp
W 187 580 ./pic/mglA1_S.bmp
W 312 580 ./pic/mglA1_S.bmp
W 437 580 ./pic/mglA1_S.bmp
W 562 580 ./pic/mglA1_S.bmp
W 687 580 ./pic/mglA1_S.bmp
W 812 580 ./pic/mglA1_S.bmp
W 937 580 ./pic/mglA1_S.bmp
; Paths of the reference and test stimuli:
P /home/subjects/exp_spk_pic/stim_ref/
Q /home/subjects/exp_spk_pic/stim_tst/
; No reference stimuli:
; R ii ii.wav 24 20
```

Table 5.1: This is an example configuration file of CoDIT for an experiment similar to that in Fig. 5.5.

N G	Switch the user interface language to 'German' instead of 'English'
U [1-65534]	Starting CoDIT from this user UID makes a dialog box appear to type-in a 5-symbol subject ID
H <text>	Permanently show a short headline
F	Show the instructions text window first
B	Switch background colour from standard grey to white
O [1-20]	AutoSave after 1 to 20 moves of a subject
V	No amplitude peak normalization
A [1-50]	Fade-in over 1 to 50ms
E [1-50]	Fade-out over 1 to 50ms
S [4000-48000]	Sample frequency between 4kHz and 48kHz
M [1024-268435456]	Pre-allocate enough memory to pre-load all stimuli (max. 512 MByte as the number is in 16 Bit words, e.g. *du -bs* is good to get an estimate of what is necessary)
X [50-2000]	X-size of the user interface in pixels
Y [50-1200]	Y-size of the user interface in pixels
L x1 y1 x2 y2	Draw line from (x1,y1) to (x2,y2). Coordinate values in the range of 0 and X/Y, respectively. Max. 64 lines
Z x y <text>	Print text at position (x,y). Max. 32 strings
R AB <filename> x y	Print a yellow button with the two-letter label AB at position (x,y). If pressed, the reference stimulus <filename> will be played. Moving is not possible
P <path>	Path to all reference stimuli. Must begin and end with '/'
Q <path>	Path to all test stimuli. Must begin and end with '/'
G <path/file>	Path of the logfile (typically ends with *.log*)
D <path/file>	Path of the instructions text file (typically ends with *.doc*)
T <path/file>	Path of the text file defining the test stimuli (typically ends with *.dat*) (see Tab. 5.3)
W x y <(path/)file>.bmp	Load an image in *.bmp* format and position it at (x,y)
; . . .	Everything after the ';' is ignored (for internal comments)

Table 5.2: Full list of all commands necessary or possible in a configuration file to define the graphical interface and its functionality.

AB <filename> x y	Print a test stimulus button with the two-letter label AB at position (x,y). If pressed, the test stimulus <filename> will be played. Moving with Click and Hold

Table 5.3: Configuration and results file for the test stimuli (typically ends with *.dat*).

5.4 Conducting an experiment

As an example, a user interface of CoDIT from an experiment by Pfitzinger (2001) is shown before it was carried out by a subject (Fig. 5.6) and afterwards (Fig. 5.7).

All test stimuli are represented as grey buttons that can be played by clicking and moved by clicking and holding. Reference stimuli are yellow and cannot be moved. In their initial position, the test stimuli are arranged in random order in the upper area of the user interface.

The subjects are seated in front of screens, wear high-quality headphones, and receive the following instructions:

Based on perceptual assessment and comparison, the stimuli are to be placed and refined on a horizontal rating scale. The height of the scale has no meaning, but is intended to make it more comfortable for the subjects to move and position.

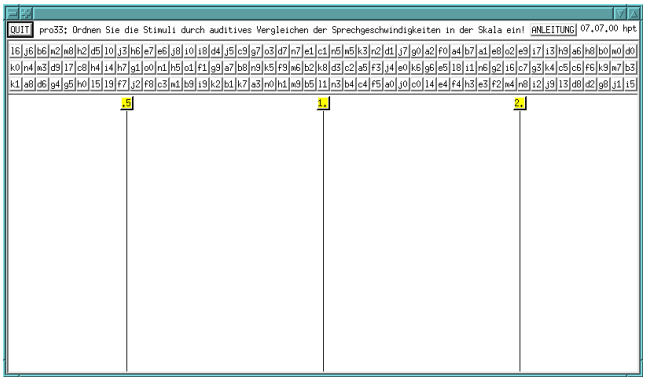

Fig. 5.6: CoDIT user interface before the start of the experiment. The stimuli to be assessed are in the upper part of the response area (cf. Fig. 5.7). (Pfitzinger 2001, p. 187)

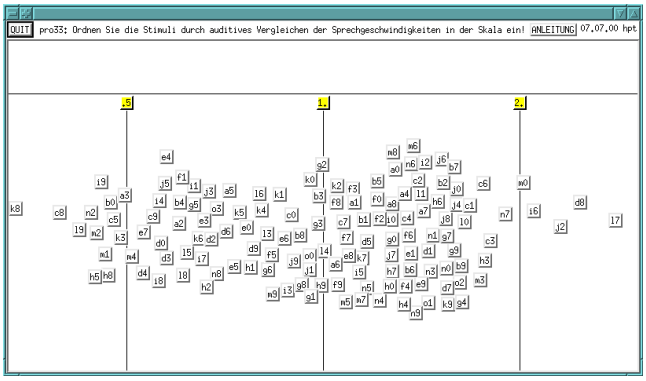

Fig. 5.7: CoDIT user interface after the end of the experiment. A subject moved the test stimuli from the upper part of the response area into the rating scale, which is spanned by the three yellow anchor stimuli with three vertical lines (cf. Fig. 5.6). (Pfitzinger 2001, p. 188)

5.5 Analysing and discussing a logfile

A striking feature of CoDIT compared to perception experiments on printed response sheets is the logging of all actions of each test subject. In order to enable processing with any software, it is a plain text file without a header, which is composed of codes for experiment-*begin* and *-end* as well as stimulus-*play* and *move* (Tab. 5.4).

On the basis of a logfile from an example subject Fig. 5.8 was drawn. It shows the course of the experiment in a way that is comparable to a punch card or perforated paper used e.g. for player pianos or crank organs. The X-axis shows the time in minutes. On the Y-axis all test stimuli from "*o3*" to "*a0*" and the three anchor stimuli ".*5*", "*1.*" and "*2.*" are arranged. Listening to a stimulus is graphically illustrated by a small point, and moving it by a small vertical line.

<time> 0 B	Time of beginning
<time> 0 E	Time of end
<time> <nr> P	Time of playing stimulus with number <nr>
<time> <nr> M x y	Time of moving stimulus <nr> to position (x,y)

Table 5.4: A logfile of a subject contains repetitions of 4 different lines. Each line starts with a time in s since the first second of 1970.

```
1230064181 0 B
1230064221 3 P
1230064222 3 M 192 124
1230064223 2 P
1230064225 2 M 171 87
1230064226 2 P
1230064228 3 P
1230064229 3 M 201 115
1230064231 3 P
1230064232 2 P
1230064242 0 E
```

Table 5.5: This example logfile shows that stimulus 3 was played 3 times and moved twice, finally to position (201,115), while stimulus 2 was played 3 times and moved once, to position (171,87). The time interval between beginning and end of the session is 61 seconds (= 1230064242 − 1230064181).

The logfiles provide new insights into the different strategies and behaviour of the participants during the course of an experiment as the following analysis and discussion of the example logfile content in Fig. 5.8 demonstrates (Pfitzinger 2001, p. 191):

The experiment took about 50 minutes. It becomes obvious how the subject listens to the stimuli starting with "*a0*" to "*a3*". In doing so, he repeatedly compares with the anchor stimuli, most frequently with "*1.*", and occasionally moves a test stimulus. In the first 22 minutes, the subject listens to all stimuli in the order of their meaningless labels and conducts a presorting. Then a 12-minute break follows.

The decrease in performance before the break is clearly observable in Fig. 5.8 by the flatter diagonal of the last stimulus block from "*k0*" to "*o3*": the angle of the diagonal corresponds to how long the test person works on one stimulus before switching to the next.

In addition to this very direct interpretation of a test protocol, remarkable statements can be made that would have been impossible with a conventional test setup:

For example, the statement can be drawn that stimuli with large standard deviations are played and moved more often than those with lower standard deviations.

In summary for all subjects of this experiment the following can be said: On average, a subject needed a total of 51 minutes, moved each stimulus 3.5 times and listened to it 10 times. Furthermore, he referred to the anchor stimulus "*.5*" 39 times, to "*1.*" 82 times, and 42 times to "*2.*".

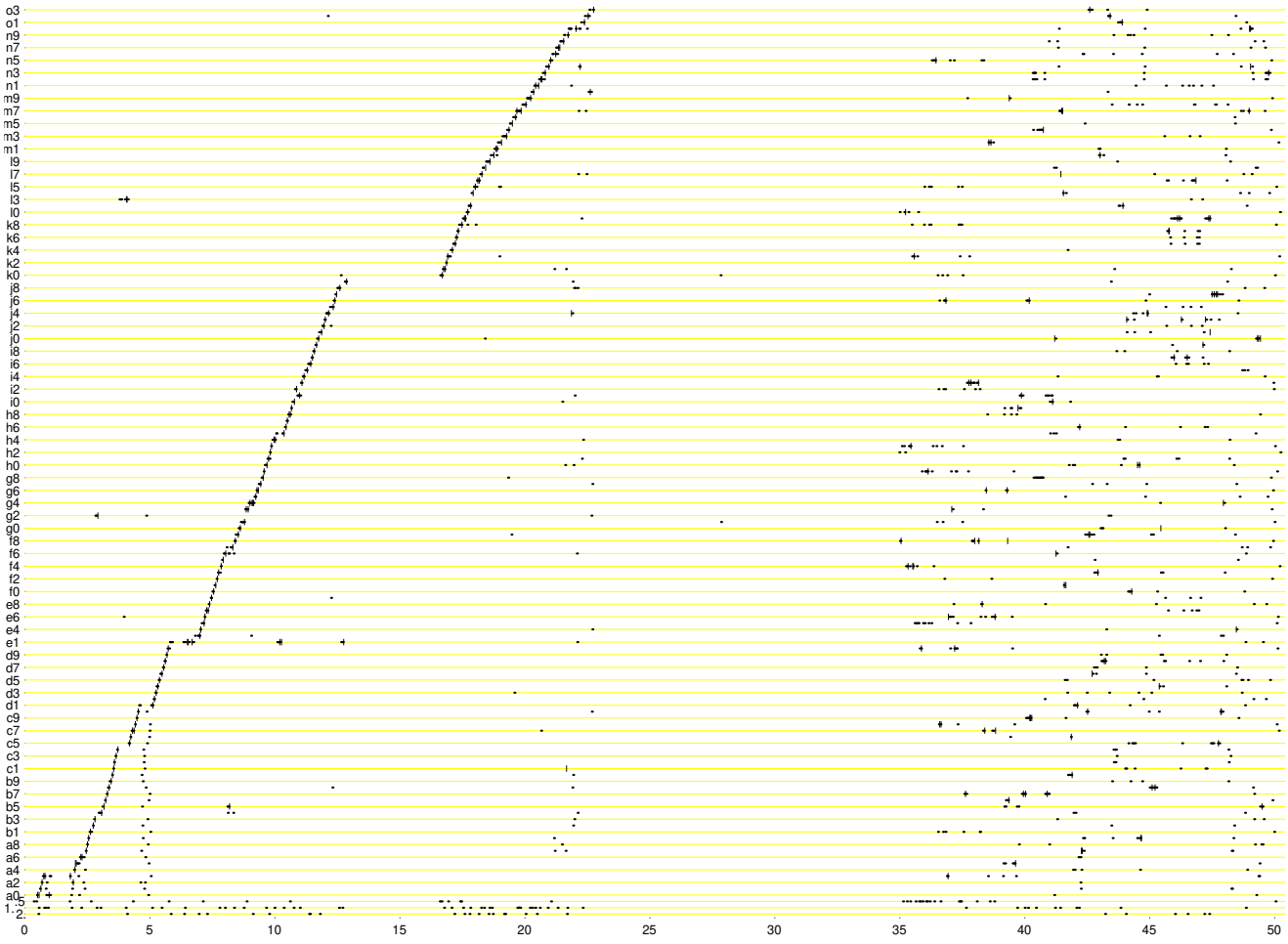

Fig. 5.8: The course of the experiment of a subject. The X-axis shows the time [min], the Y-axis the stimuli (Pfitzinger 2001, p. 191).

5.6 Discussion

Perception tests with two related answer scales extraordinarily profit from the freely shaped two-dimensional response area of CoDIT. Basically, this perception test tool combines identification and discrimination tests, as the participants are enabled on the one hand to evaluate the stimuli absolutely and position them accordingly along the spatial dimensions of the response area, but on the other hand they can always compare. Thus, its usage is preferable if perceptual data shall serve as a reference in modelling or, more generally, if higher precision and accuracy of judgements is desired, which is a result of CoDIT's possibility to play, compare, and reposition stimuli repeatedly.

But if a clocked listening test with a constant rating time interval for each stimulus is necessary or if the participant is not to compare different test stimuli with each other, then CoDIT is not recommended.

Obviously, the researcher is free to design nominal (e.g. Fig. 5.5), ordinal, interval, or ratio scales with the drawing commands of CoDIT. It is his responsibility to decode the final perception data in accordance with the original scales, and to not over-interpret the precision of the resulting coordinates.

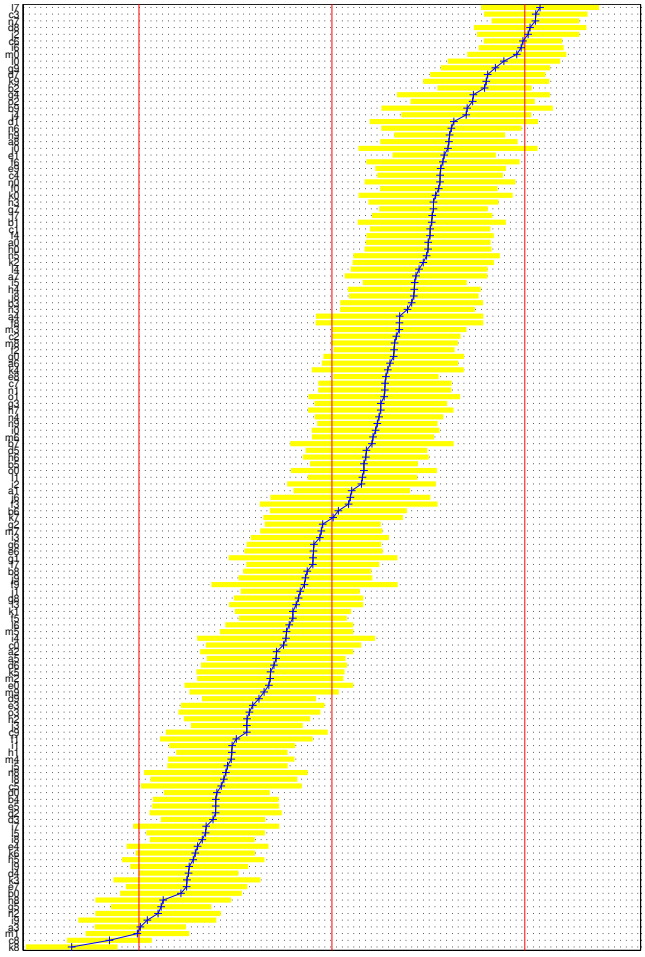

Fig. 5.9: The rating scale leads from left to right with the three anchor stimuli at the vertical lines. The 141 test stimuli are vertically ordered according to their mean judgements (See Fig. 5.7).

An important question is if a scale is actually an interval scale, when it was graphically designed so that it looks like it and when the participants were instructed to place the stimuli along the scale and were also explicitly asked to ensure that perceived differences corresponded to distances on the scale (e.g. Fig. 5.7).

It can be assumed that the subjects can deal with the proportions of the area on an interval-scaled basis if it were only a purely graphical task. However, previous results have shown that anchor stimuli can distort the interval scale: It is locally stretched in the immediate proximity of an anchor stimulus because the smallest audible differences lead to clearly different positioning (Fig. 5.9).

In contrast, stimulus positions in the middle between two anchor stimuli tend to be compressed, since there the audible difference to the two anchor stimuli is always very big, but only very small to neighbouring test stimuli. Of course, distortions in the proximity of anchor stimuli could be mathematically compensated e.g. using the arctangent function. But (1) there is not yet a sufficient foundation for this step, and (2) the deviations are smaller than the inter-subject variation. The latter means that the effects of a compensation, e.g. on modelling, would probably only be very small.

But how strong the distortions of the interval scale are depends on the amount and careful selection of the anchor stimuli. A preliminary experiment could serve to find further anchor stimuli at approximately equidistant intervals. The reallocation of these test stimuli to additional anchor stimuli would then lead to more uniform results in a main experiment. Anyway, the consistency of the results is by no means the ultimate goal of perception experiments. In addition, what has just been said under (2) applies.

5.7 Conclusion

During the last 12 years CoDIT was applied in 16 perception experiments with more than 400 subjects in total. According to the experiences of various researchers as well as participants, it is admissible to summarize that the tool is absolutely reliable, lightweight software, and easy to use. CoDIT together with working example files can be freely obtained from the author via email.

Since CoDIT offers subjects the possibility of discrimination in addition to identification, it was helpful to apply it in experiments that aimed at investigating L1 and L2 differences and effects. It was already applied in analysing L1 effects on vowel perception (Dioubina & Pfitzinger 2002), assessing language learning success with regard to word stress (Bissiri et al. 2006; Bissiri & Pfitzinger 2009) (see Fig. 5.10), and intercultural differences between German and Japanese listeners in speech rate perception (Pfitzinger & Tamashima 2006).

A valuable experience was that more than 150 stimuli for assessment exhaust the subjects in a way that they ex-

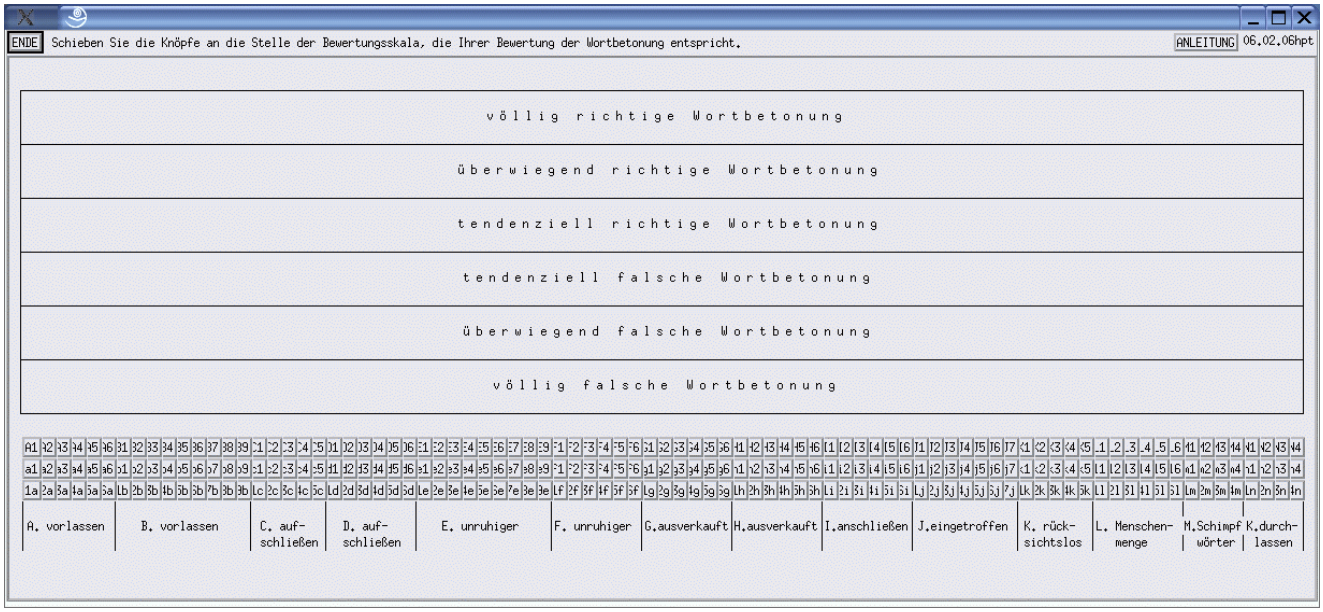

Fig. 5.10: "User interface for the word stress correctness assessment test. [...] The test was implemented with the tool CoDIT." (Bissiri et al. 2006; Bissiri & Pfitzinger 2009)

pressly prefer not to participate again in a similar perception test. Therefore, in cases where assessments of more than 150 stimuli are required, it is recommended that the subject be allowed to complete the task in more than one day, with breaks as often as desired. In general, judging 80–120 stimuli was found to be a feasible task, albeit still time-consuming.

What remains to be done is (1) the introduction of a third rating dimension, e.g. in form of changing the button colour on a *colour-temperature-* or *black–white*-scale or changing the size or shape of the buttons, (2) a fast replay of the participant's actions as a visual animation with optional focus on a short list of interesting stimuli to inspect and understand user behaviour in detail, and (3) a supervisor tool that permanently reads the logfile (which during a course of an experiment has a continuously increasing file size due to the *AutoSave* option) and displays all stimulus playbacks and shifts in real-time or even performs some real-time analysis.

Bibliography

Bissiri, M. P. & Pfitzinger, H. R. (2009). Italian speakers learn lexical stress of German morphologically complex words. *Speech Communication 51*(10), 933–947.

Bissiri, M. P., Pfitzinger, H. R. & Tillmann, H. G. (2006). Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis. In: *Proc. of the 11th Australian Int. Conf. on Speech Science and Technology (SST '06)*. Auckland; New Zealand, 24–29.

Boersma, P. & van Heuven, V. (2001). Speak and unSpeak with PRAAT. *Glot International 5*(9/10), 341–347.

Dioubina, O. I. & Pfitzinger, H. R. (2002). An IPA vowel diagram approach to analysing L1 effects on vowel production and perception. In: *Proc. of ICSLP '02*, vol. 4. Denver, 2265–2268.

Hoole, P., Gfroerer, S. & Tillmann, H. G. (1990). Electromagnetic articulography as a tool in the study of lingual coarticulation. Forschungsberichte (FIPKM) 28, IPSK, Univ. München, 107–122.

Pfitzinger, H. R. (1994). Vokalperzeption und akustische Merkmale. Master's thesis, IPSK, Univ. München.

Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In: *Proc. of ICSLP '98*, vol. 3. Sydney, 1087–1090.

Pfitzinger, H. R. (1999). Local speech rate perception in German speech. In: *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 2. San Francisco, 893–896.

Pfitzinger, H. R. (2001). Phonetische Analyse der Sprechgeschwindigkeit. Forschungsberichte (FIPKM) 38, IPSK, Univ. München, 117–264.

Pfitzinger, H. R. (2003). Acoustic correlates of the IPA vowel diagram. In: *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 2. Barcelona, 1441–1444.

Pfitzinger, H. R. & Tamashima, M. (2006). Comparing perceptual local speech rate of German and Japanese speech. In: *Proc. of the 3rd Int. Conf. on Speech Prosody*, vol. 1. Dresden, 105–108.

Thon, W. (1982). Microprocessor-controlled reaction-time measurement. Arbeitsberichte (AIPUK) 18, IPDS, Univ. Kiel, 84–91.

Tillmann, H. G. & Pfitzinger, H. R. (2004). Applying the Munich Parametric High Definition (PHD) speech synthesis system to the problem of teaching Chinese tones to L1-speakers of German. In: *Proc. of the Int. Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages (TAL)*. Beijing, 185–188.

Willing, M. (2010). Relationships between auditorily and visually evaluated speech characteristics. Arbeitsberichte (AIPUK) 38, IPDS, Univ. Kiel, 35–42.

Zierdt, A., Tillmann, H. G. & Hoole, P. (1996). Towards a three-dimensional articulographic system. *J. of the Acoustical Society of America 100*(4, Pt. 2), 2662.