DUTCH ESPRIT PROJECT 2589
(SAM)
MULTI-LINGUAL SPEECH INPUT/OUTPUT
ASSESSMENT, METHODOLOGY AND STANDARDISATION
3 March 1992

Introduction

This report provides the documentation for the Dutch recordings that were made during the course of the ESPRIT-SAM effort to collate the EUROM.1 speech database. The recordings took place in October 1990 at the TNO/IZF facilities in Soesterberg, in the Netherlands.

The Dutch EUROM.1 database consists of three corpora:

- the many talker corpus.

- the few talker corpus and

- the very few talker corpus.

These corpora each contain material which is comparable to similar corpora within the databases for the other languages for which EUROM.1 has been gathered.

The recording of the entire database was controlled by version 4.0 of the EUROPEC programme, which has been specifically developed within the ESPRIT-SAM project to support the recording of EUROM.1. The recording protocols that specify the way the EUROPEC programme should be used within this recording are described in SAM-RSRE-15.

The actual recordings were made by three operators, Wim van Golstein Brouwers, Aad Belt and Willem Roel, all affiliated with PTT Research. The recordings took place in the anechoic chamber of TNO/IZF in Soesterberg. The passages and sentences were defined by Cyriel Houben. The phonotypical and broad phonetic transcriptions were made by Leonie Bos.

1 Database contents

A total of 64 subjects were selected for the database recordings. These 64 subjects constituted the many talker set. Ten of these subjects contributed extra material and thus made up the few talker set. Four of these subjects contributed even more, providing the very few talker corpus.

1.1 Prompting texts

The EUROM.1 prompting texts consist of several parts.

- passages
- sentences
- numbers
- CVC's
- CVC's in carrier phrases

The orthographic transcriptions for each of these prompting texts are given in appendix A.

1.1.1 Passages

In total, a set of 40 passages has been constructed. These passages each consist of a block of 5 related sentences on topics of a very general nature. The prompting texts for the passages were presented as one single block, so that subjects were able to produce the text as a single entity.

1.1.2 Sentences

A set of 10 blocks of 5 sentences each was provided, in order to balance the uneven phoneme distribution in the passages. The sentences in this set were not related and were presented one at a time.

1.1.3 Numbers

100 numbers were solicited in 5 sets of 20 numbers each. These numbers were between 0 and 10000. The pronunciation for these numbers was provided between brackets in orthographic form, thus eliminating the risk of alternative pronunciations (e.g. fifteen hundred rather than one thousand five hundred).

1.1.4 CVC's

A total of 66 different CVC words (also containing a few CCVC words) were collected. These 66 CVC's were split into 2 groups of 33 CVC's each. All of the 10 subjects in the few talker group pronounced these CVC's five times each.

1.1.5 CVC's in carrier phrases

The CVC words were embedded in a phrase consisting of one word before and one word after the CVC word. Five different word pairs were used:

- lees ... openbaar
- regel ... stop
- evenzo ... legaal
- attentie ... kapot
- vrolijk ... iedereen

A total of 10 lists of 33 such phrases in each were compiled. These lists also included 5 repetitions of all 10 context words that were used. All 4 of the very few subjects pronounced all 10 lists.

1.2 Recorded corpora

In total, 3 different corpora were compiled. They were the Many Talker, the Few Talker en the Very Few Talker Corpora.

1.2.1 Many Talker Corpus

The Many Talker Corpus was collected from 64 subjects, 30 of whom were male and 34 were female. Each of the 64 subjects in the Many Talker Corpus spoke the following material:

- 3 passages
- 1 block of 5 sentences
- 5 blocks of 20 numbers

1.2.2 Few Talker Corpus

The Few Talker Corpus was collected from 10 subjects, all of whom had acted as a subject in the Many Talker Corpus. Of these 10 subjects, 5 were male and 5 were female. Each of the 10 subjects in the Few Talker Corpus spoke the following material:

- 5 times 2 blocks of 33 CVC's
- 5 times 5 blocks of 20 numbers
- 5 blocks of 5 sentences
- 15 passages

1.2.3 Very Few Talker Corpus

The Very Few Talker Corpus was collected from 4 subjects, all of which had acted as a subject for the Few Talker Corpus. 2 of these subjects were female, 2 were male. Each of the 4 subjects in the Very Few Talker Corpus spoke the following material:

- 10 blocks of 33 CVC phrases

1.3 Summary of recorded corpora

Many Talker Corpus (64 subjects):

- 3 passages, each 5 sentences long
- 1 block of 5 unrelated sentences
- 100 numbers between 0 and 10000

Few Talker Corpus (10 subjects):

- 15 passages, each 5 sentences long
- 5 blocks of 5 unrelated sentences
- 5 repetitions of 100 numbers between 0 and 10000
- 5 repetitions of 66 CVC words

Very Few Talker Corpus (4 subjects):

- 66 CVC words embedded in 5 different carrier phrases
- 5 repetitions of the 10 carrier words

2 Subject selection

The subjects were selected to ensure a wide variety, according to the guidelines given in Appendix R8, (Ref 3). They were primarily chosen from the TNO/IZF staff, with some PTT Research employees, complemented with 12 female students who were selected in order to even the balance between male and female subjects.

The subjects each filled out a subject inquiry form, listed in appendix B. The contents of these forms were compiled in a subject database "speakers.dbf", according to the EUROPEC standard.

The distribution of the subjects across various age categories can be determined on the basis of the speaker database. Table 1 presents this distribution.

Age Category	**Talker ID (male)**	**Talker ID (female)**
16-25	2 [LO],MJ	15 AR,DK,[BS],ED,

		GM,HA,HS,KH, OL,PM,PS,{[SD]}, {[SS]},VM,WG
26-35	11 [BW],{[DH]},EA,GF, GW,MF,MM,SW, LM,LR,VK	7 HM,KG,KO,{[LJ]}, VA,VJ,VN
36-45	9 AJ,BA,BR,DJ, AL,BE,BL,FE,VO	8 GR,KJ,RH,RJ, HP,PA,PN,VC
46-55	8 AE,BJ,PL,PP, [RW],SH,SL,[VS]	2 JJ,[VE]
56-	0	2 BN,GH

Table 1: Distribution of subjects over the different age categories.

Few talkers are marked with [], while talkers from the Very Few Talker group are marked with {}.

2.1 Distribution of subjects over prompting texts

For both the Many Talker Corpus and for the Few Talker Corpus, part of the material was produced by all subjects, while another part was produced only by a certain fraction of the subjects for that corpus. The selection of texts was such that each subject had a unique selection of material to produce. In appendix C.1, the texts for each subject and the subjects for each text in the Many Talker Corpus are listed. In appendix C.2, a similar listing is presented for the Few Talker Corpus, while appendix C.3 contains these listings for the Very Few Talker Corpus.

3 Recording protocol

3.1 Recording environment

The recordings took place in an anechoic room. The subject could read the prompts from a monitor, which was positioned 80 cm from the subject's lips, at a slight angle (10 degrees), to minimize reflections from the monitor to the speaker or microphone.

In order to minimize disturbing sounds, fluorescent lights and air conditioning were switched off. All persons working in adjacent rooms were asked to be quiet, although the sounds that they would have made could not be heard by ear. Reflections off the table in the anechoic room were minimized by the use of a table cover of sound absorbing material.

3.2 Recording equipment

The following recording equipment was used for the recordings:

- a B&K half inch microphone (4165)
- a B&K digital sound level meter (2230) which also acted as a microphone amplifier
- OROS AU21 A/D board
- Luxman DAT recorder
- a laryngograph

The AC output of the sound level meter (SLM) was connected to both the line input of an OROS AU21 board and the left channel of a Luxman DAT recorder. The signal was split with a (TNO-IZF) self-made signal splitter in such a way, that no degradation to the audio signal was introduced. The laryngograph was used for recordings of the CVC words and the CVC phrases in the Few Talker Corpus, with the exception of the recordings of one female subject. The sensors were positioned on either side of the thyroid cartilage. The narrow-band output (100-550 Hz) was connected to the right channel of the DAT recorder. All of this equipment was situated in the adjacent room, with the obvious exception of the microphone itself.

During the entire session, the DAT recorder was kept running.

The settings of the OROS board were taken from the recommendation. This results in a 20.000 Hz sample frequency, the use of the line input and variable input gain. The oversampling, however, was not set to 4 but to 2, because the OROS hardware used will not allow for a setting of 4 at a sampling frequency of 20 KHz.

The speech signal was recorded on hard disk and with frequent intervals, this disk was backed up and erased. The backup medium was a Maxtor TAHITI magneto optical disk drive.

3.3 Recording mode and prompting style

The recordings contain no speaking errors. The inter-utterance pauses are captured in their entirety and there are no discontinuities within a take. When a speaking error or a software error occurred, the entire take was re-recorded.

A "mixed" timing strategy was used, which means that the prompt was on the monitor for a period that was controlled by a predetermined duration and the endpoint of the utterance. This means that if the subject is slow, the prompting system will also slow down. If, however, the subject is extremely fast in responding, the recording system will not follow suit. The length of the predetermined interval is text-dependent. For CVC words and short numbers, the duration is 2 s.; for CVC phrases and medium length numbers, it is 3 s.; for long numbers, it is 4 s.; for sentences, this duration is 7 s. and for passages, the length is 25 s. (shortest passage), 30 s., 35 s. or 40 s. (longest passages).

The recordings were made in the continuous mode, except for the calibration recordings, which were made in manual mode.

The relevant parameters for the recordings in continuous mode are as follows:

- Triggering level:	-30 dB
- Extinction level:	-30 dB
- End silence:	1000 ms. for CVC'S, CVC phrases and numbers 2000 ms. for sentences and passages
- Signal head:	200 ms.
- Signal queue:	500 ms. for CVC'S, CVC phrases and numbers 1500 ms. for sentences and passages

Explanation of terms

- triggering level: The level that the signal must reach, to trigger the recording process.

- extinction level: The level that the signal must not reach, to signal a piece of silence.

- end silence: The minimal duration of silence that determines the end of a recording.

- signal head: The amount of silence before the first utterance that is to be recorded onto disk.

- signal queue: The amount of silence after the last utterance that is to be recorded onto disk.

3.4 Recording control

The subject was prompted by the EUROPEC programme, running on the SESAM workstation, positioned outside the anechoic room. The EUROPEC system was controlled by a recording manager (operator), viewing the standard SESAM monitor using the EGA graphics mode. For all recordings, the subject was able to see the prompted text and a level meter without stress on a second standard SESAM monitor inside the anechoic room, also connected to the video output of the workstation. The monitor was positioned in such a manner, that the acoustical and electrical pickup by the microphone is not audible without amplification.

The operator was able to speak to the talker in the anechoic room through a two-way intercom system and could also see the talker on a video screen. The gain of the intercom was set to a level at the ear, representing speech at about 1 m. from the lips of an average talker. During the recordings, the microphone was not able to pick up any sounds from the intercom. The operator was continuously listening to the talker, by monitoring the headphone output of the DAT recorder. This allowed the operator a 100% check on the speech material contents. To reduce fatigue on the part of the operator, two operators took turns.

During the test and actual recordings, speaking effort was controlled by the operator and by the subject. This was done by using a speech level meter, displayed on the prompting monitor by EUROPEC. The microphone amplifier was set so that the normal peak level of the speech reached a reference point, 10-15 dB below peak. The subject was controlled for consistent speaking level during the subsequent takes by both the operator and himself (he was asked to keep to the reference point as much as possible).

The total amount of material recorded in each session was restricted to 14 minutes, representing approximately a 40 minute elapsed time session. This limit minimizes the stress on the subject and represents the maximum time that could be recorded on the SESAM hard disk. The Few Talker Corpus was divided into 4 sessions, so that the talker could take a break and the recordings could be stored on the backup medium.

The operator tried to make the subject feel at ease. However, no attempt was made to control any speaking effect due to the time of day of the session.

3.5 Recording procedure

All subjects took part in the recordings having made an appointment. Arrangements were made so that the TNO-IZF and PTT Research employees were able to make the recordings during working hours. The students were paid for their time and arranged to come in for a morning or an afternoon.

Before each session, the operator changed the batteries of the sound level meter to ensure that the sound level meter worked properly. He also checked the recording station and apparatus. At the beginning of each session, the recording operator welcomed the subject and asked them to fill out a questionnaire (appendix B). Next, the operator provided a written outline of the procedure for the recording session. This briefing (appendix D) was provided in written form, in order to ensure consistency between sessions. The subject was also made familiar to the speech material by reading some examples beforehand.

Before the recordings actually started, the operator entered the subject's code and characteristics into the subject database. The subject was then taken to the anechoic room where he or she sat in the chair. The microphone was placed in the correct position and the distance from the microphone to the subject's lips was measured to be 50 cm. The subject was then asked to read some text from the screen for calibration so that the operator could read the peak level from the sound level meter (SLM). If this peak level was lower than 75 dB, the SLM was set to 80 dB full scale display (FSD), otherwise to 90 dB FSD. The gain in the EUROPEC programme was 0 dB, ensuring that the peak level is about 12 dB below the maximum possible recording level. This level is safe and reduces the possibility of takes having to be rerecorded due to overload, which was thought to reduce stress on the (naive) subjects.

The operator checked if the fluorescent lights and air conditioning were switched off. The correct file (CO80DB.TXT or C09ODB.TXT) was copied to CO.TXT and the recordings on the DAT recorder were started. First, 10 s. of a 1 KHz calibration (B&K calibrator 4230) signal was recorded with FSD = 100 dB.

Before each recording of a list for the database, the subject could first practice speaking some typical material. For the passages, the subject is given the opportunity to preview all the texts that had to be spoken.

During the recordings, the operator monitored the speech production with reference to the prompted text. If there was a speaking error, a software error or an overload error, the take was stopped and the subject was asked to start the relevant take from the top. The disk file was discarded and the same take number was re-used for this new recording. The DAT recorder was never stopped during a session.

The EUROPEC software provides a function for the validation of the material that has been recorded. This feature only functions reasonably well, if there have been no software errors. Since such errors occurs fairly frequently, this validation should really be done immediately. This, however, took so much time (the same amount of time as the actual recordings themselves), that in practice, this validation was only performed on a small fraction of the recordings. Afterwards, a separate tool was used to validate the entire corpus.

After all recordings were made, the DAT recorder was stopped and the subject thanked for his or her cooperation. The operator then copied the material to TAHITI optical disk and deleted the files from hard disk.

The recordings of the speech material for the Many Talker Corpus took about 60 min. for each subject. The introduction of the subject took about 5 min., the calibration 5 min., the speech recordings about 40 min. (including 8 min. for the recording of some additional, non-ESPRIT material) and the copying of the recorded material to optical disk took 10 minutes. For each subject, this resulted in some 14 min. (or 30 MBytes) worth of speech stored on disk after a session.

The recordings of the speech material for the Few Talker Corpus took about 130 min. for every subject. The introduction took about 5 min., the calibration 5 min., the speech recordings about 4 times 20 min. and copying the data to optical disk took 4 times 10 minutes. For each subject, there was about 50 min. (or 110 MBytes) of speech recorded on disk after a session.

The recordings of the CVC phrases in the Very Few Talker Corpus took about 60 min. for each of the 4 subjects. The speech recordings took 2 times 20 min., while copying the files to optical disk took 2 times 10 minutes. For each subject there was about 28 minutes (or 60 MBytes) of speech recorded on disk.

With the subjects in the Few Talker Corpus, the first material to be recorded were the CVC's. These recordings were with the laryngograph running, the output of which was being recorded on the right channel of the DAT. With the subjects in the Very Few Talker Corpus, the CVC phrases were recorded immediately after the CVC recordings. The laryngograph could thus remain in place.

3.6 Integrity checks

After recording a take, the quality and the item end-point labelling of the recordings were checked in some cases. This was not done for all of the material, because of the time consuming nature of this check. At a later date, all of the material was checked.

3.7 Calibration

The recording chain was calibrated in three ways.

1) A B&K 1KHZ 4230 calibrator was placed over the microphone and 10 s. of this tone was recorded before every session.

2) Every week, a more extensive set of calibration signals was recorded.

2a) The B&K 1KHZ 4230 calibrator was placed over the microphone and 10 s. of this tone was recorded with the SLM on 100 dB FSD.

2b) A rectangular wave with 4:7 mark-space ratio of 20 Hz and 100 mV peak-peak was connected to the microphone input (a B&K input adaptor JJ 2614 replaces the microphone) and 10s. of this signal was recorded with the SLM on 100 dB FSD.

2c) The line input was terminated by a 50 Ohm impedance near the SLM and 10s. of silence was recorded.

2d) The microphone was placed in its normal position, no subject was involved and 10 s. of silence was recorded with the SLM on 70 dB FSD.

3) Two times (in the beginning and at the end of the recordings), the anechoic room conditions were assessed by the bursts of 6 balloons. All balloons were orally inflated to a 270 mm. diameter, measured by a U-shaped board. The pressure of each balloon was measured by a water manometer. The pressure of the balloons was (in order of burst) 27.2, 25.5, 26.0, 23.4, 25.4 and 28.5 cm. water pressure for the first session and 23.5, 21.2, 27.5, 31.5, 24.1, 32.0, 27.5 cm. water pressure for the second session. For the second session, 7 balloons were burst. The temperature was 20°C for the first session and 18°C for the second session. The ambient pressure was 1019 mbar for the first and 996 mbar for the second session. The recordings were made with the chair removed and using the standard recording setup with the balloon in the position of the lips. The distance of the microphone to the balloon was 50 cm. When making

the recordings, the balloons were held at arm's length and burst with a sharp object. The bursts were recorded with the SLM on 120dB FSD.

3.8 Laryngograph recordings

The original specifications for the EUROM.1 recordings were such, that part of the material should also be recorded with a laryngograph. These recordings, however, would not be included in the database itself, but would be available on digital audio tape (DAT). A later revision of the specifications made only after the Dutch recordings were completed, implied that this laryngograph material should also be stored on disk and be included in the database itself. In order to accommodate these revised specifications, the material was extracted from DAT in digital form, using the OROS AI interface. This signal was then downsampled to 20.1 KHz and stored on disk. This stereo signal was then split into a microphone and a laryngograph signal. This signal will be referred to as the redigitized signal. The microphone signal differs from the original digitized version of that same take in 2 aspects.

- The sample frequency of the redigitized recording is 20.1 KHz instead of 20 KHz.
- The amount of silence preceding the first token in the redigitized version will differ from the silence preceding the same token on the original sampled data file.

3.9 Collation of recordings

The recorded material has been collated and is available on DAT and on EXABYTE tapes. It will be stored on a single large capacity WORM as well as on a number of MOD magneto optical disks shortly.

The files were named by the recording software EUROPEC according to the following naming scheme.

```
TTPPXXXX    CNF
|  |  |       | | |
|  |  |       | | ---  Filetype          S: Sampled speech
|  |  |       | |                        O: Orthographic file
|  |  |       | |
|  |  |       | -----  Nationality       H: Dutch
|  |  |       |
|  |  |       -------  Corpus index      N: Numbers
|  |  |                                  S: Sentences
|  |  |                                  P: Passages
|  |  |                                  W: CVC words
|  |  |                                  C: Calibration
|  |  |
|  |  --------------------  Serial number (one per take)
|  |
|  -----------------------  Prompting text name
```

```
 |
---------------------------- Talker ID
```

An exception to these naming strategy is made for the recordings that were redigitized from DAT. They have the same filename root (i.e. TTPPXXXX) but have the extension ".?H2" for the microphone signal and ".?HL" for the laryngograph signal. There are no label files directly associated with these recordings.

3.10 Data quantity of the Dutch EUROM.1 database

The following table shows the data quantity for the different parts of the three corpora. The material that was rerecorded from DAT is denoted as "redigitized".

	Numbers	**Phrases**	**Passages**	**cvc**	**cvc-phrases**	**Total**
Many						
1 subject	12.3	1.5	3.6			17.4
64 subjects	787	93	232			1112
Few						
1 subject	59.2	7.3	18.5	28.5		113.5
10 subjects	592	73	185	285		1135
Very few						
1 subject	64.5					64.5
4 subjects	258					258
Redigitized				600	520	1120
Total	1379	166	417	885	778	3625

Table 2: distribution of data across the different corpora within EUROM.1

4 Conclusion

Recording EUROM.1 was a tremendous amount of work. Careful planning made it possible to complete the entire recording procedure in just over three weeks. It was, however, considered necessary to have 2 recording operators present full time, due to fatigue.

The subjects themselves also risked some fatigue. Some of them were involved with recording for up to 3 hours (including breaks). Some subjects drank coffee during these pauses. This occasionally caused stomach rumbles during the recordings. After noticing this, the operators asked the subjects to refrain from drinking coffee until after the recordings had been completed.

The dual presentation of numbers (both as a digit string as written out in full) caused some remarks from the subjects. It was suggested that it would be clearer, if the two representations were presented on separate lines on the prompt screen.

The prompt screen was thought to produce some background noise. Consequently, it might be better to use an LCD screen in future.

During a validation procedure that was followed after the database had been collected, it was found that the beginning and ending of weak fricatives had sometimes been "missed" by the end-point detection of the EUROPEC software. On account of the signal head and signal queue, some surrounding silence was also recorded. The missed parts of speech were found to be thus automatically included.

5 Bibliography

SAM (1990) ESPRIT project 2589 (SAM) Multi-lingual speech input/output assessment methodology and standardisation, Interim report year one, ref: SAM-UCL-GO2