# CONTROLLED ELICITATION AND PROCESSING OF SPONTANEOUS SPEECH IN VERBMOBIL

*M. Pätzold, M. Scheffers, A. Simpson and W. Thon*
*IPDS, Kiel, Germany*

## ABSTRACT

Within the framework of the German Verbmobil project, a large amount of spontaneous dialogue data had to be collected. This paper describes the recording environment and the means of elicitation and transcription which have been developed at Kiel to fulfill this task.

## INTRODUCTION

The ultimate goal of the Verbmobil project [1] is the development of a portable translation system with voice input and output.

The aim of data collection in the first phase of Verbmobil was to provide a large amount of German spontaneous dialogues associated with appointment making. The data should consist of high quality speech signals together with their orthographic and phonemic transcriptions. Part of the signals should also be segmented and labelled. The dialogues should be elicited in a controlled situation, but still be as spontaneous as possible.

Part I of this paper describes the technical details of the recording environment and the signal processing developed to meet the requirements imposed on the speech recordings. Part II describes the elicitation of appropriate material and the subsequent steps involved in its transcription and segmentation.

## PART I: SIGNAL RECORDING AND PROCESSING

The following requirements were imposed on the speech recordings:

- Synchronous capture of the speech signals of two dialogue partners.
- High quality recording (low background noise level, large dynamic range).
- The actually recorded "turns" should not overlap in time.

We furthermore needed to reckon with a recording session lasting up to an hour. The end product should be a series of speech files each containing one "turn", arrived at with as little manual labour as possible.

### Recording Environment

To meet these requirements, a hardware/software recording environment has been implemented with the following features (see Figure 1):

- The dialogue partners are placed in separate sound-treated rooms. They communicate via headsets.
- The speech signals are recorded directly to hard disk into a multiplex stereo file (2x16bit/16kHz), on a PC AT486/66 platform with about 500MB disk space, sufficient for recording sessions in excess of one hour.
- Both microphone signals are recorded on DAT for backup purposes.
- A DSP Motorola 96002 controls the high quality analogue I/O channels as well as the digital I/O.
- The dialogue is controlled by speak-buttons and lights. Both speakers may request their input channel by pressing their button. Requests are granted on a "first come, first served" basis. Service is indicated to a speaker by his light being turned on. Thus, only one speaker's signal is recorded at a time.
- The experimenter may at any time during the recording session communicate with the dialogue partners via his microphone without interfering with the signal recording.

The DSP programme controls the A/D and D/A conversion and monitors the button actions.

If no buttons are pressed, a zero signal is output to both headphones, a distinct constant marker signal is recorded on both channels and both lights are turned off.

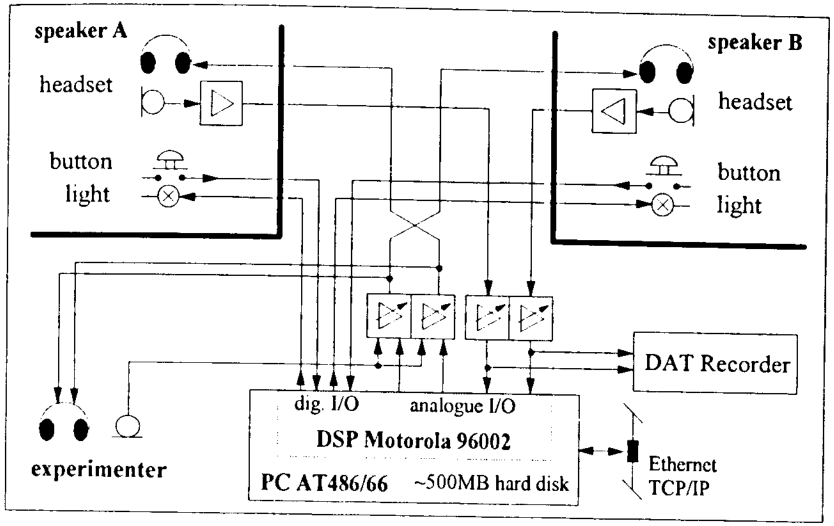

*Figure 1. Recording environment: hardware set-up.*

As soon as the first speaker presses his button, his speech signal is routed to his partner's headphones and recorded on his channel, his light is turned on until he releases his button.

If, during this time, the other speaker presses his button, his light remains turned off, a zero signal is still output to his partner's headphones, but a different constant marker signal is recorded on his channel.

Thus, the marker signals recorded on both channels reflect the exact timing of button actions. Furthermore, the DSP programme continuously checks the input sample levels and signals these to the PC programme.

Running in parallel, the PC programme does the actual data transfer to the hard disk and provides the user interface to the experimenter. During a test session a peak level display may be used to optimize the recording level for the extreme dynamic range of spontaneous utterances.

### Signal Processing

The resulting data file is transferred via Ethernet to a UNIX workstation, where it is de-multiplexed and split into two files, one for each channel. At the same time the embedded markers are detected and converted to a list with the time intervals of the turns. After inspection and, if desired, correction of the interval markers (e.g. because a speaker has released his button for a short time within his turn), a second programme uses them to split the two files into the desired series of turn files. Starting from the original multiplexed file, the names of the respective output files are generated automatically and contain at the end stage a code for the dialogue scenario (e.g. appointment making), a code for the recording site, a recording identification number and information on the channel (speaker) and the position of the turn in the dialogue. Finally, a programme is available to convert the files from local format to the delivery format.

## PART II: ELICITATION, TRANSCRIPTION AND SEGMENTATION

In this section we describe the elicitation of appropriate material and the subsequent steps involved in its transcription and segmentation.

### Elicitation

The recordings had to contain the following material:

- names of months
- dates
- names of days
- names of holidays
- times
- deictic time expressions
- proper names
- names of towns
- spelling

To guarantee systematic coverage of the material the following scenario was developed [2].

Each speaker was given a set of calendar sheets each covering a two-month period, together with timetables covering the weekdays. The calendar sheets and timetables were placed face down in a pile in front of the speaker together with a pen for making notes.

Apart from the names of months, dates and the names of days, the calendar sheets also contained names of holidays, exemption blocks (shaded areas representing days on which the speaker could not make an appointment) and simple appointments. The timetables had the names of days, times and exemption blocks.

The calendar sheets and timetables served to elicit the names of months, the names of days, times and the names of holidays. Appointment entries in the calendar sheets were designed to elicit names of German towns, e.g. "Dienstreise nach Kiel" ("Business trip to Kiel").

In order to make speakers utter letter names, appointments had to be arranged at an exhibition ("IAA in Frankfurt") and at a conference ("ICPhS in Stockholm").

Finally, deictic time expressions were elicited using a portion of a timetable. The names of the days were left out and the speakers were told that the first day on the timetable was today, and that three meetings had to be arranged over the next two days, i.e. today, tomorrow and the day after tomorrow.

Each recording session was split up into eight tasks. Each task involved the speakers arranging three appointments in the period specified on a calendar sheet. The appointments were noted at the bottom of the calendar sheet and also briefly explained by the experimenter.

The first seven tasks allowed the twelve months of the year to be covered with six calendar sheets. The first task was used as a dummy to get the speakers accustomed to the set-up and enable recording levels to be set. The eighth task involved the elicitation of deictic time expressions using the cut-down timetable.

Before the first recording, the speakers were instructed on the tasks and on the use of the speak-button.

### Orthographic Transcription

The transcription system provides an orthographic representation of the dialogues [3]. The system must fulfill two requirements. First, it must be simple to allow for a relatively fast transcription of a large amount of data. Second, it must attempt to meet the demands of both signal processing and linguistics.

As well as transcribing the lexical content, the system must also capture characteristic aspects of spontaneous speech.

For lexical items and semantic-syntactic structure the transcription is based on the Duden conventions [4] as far as possible. In addition, the following objects typical for spontaneous speech are included in the system:

- interjections
- agreement and negation particles
- particles indicating hesitation
- non-words (neologisms, slips of the tongue)
- laughing, coughing, lip-smacking etc.
- articulatory lengthening
- breathing and pauses
- breaks and repairs
- stretches of utterance, either poorly understood or not understood at all by the transcriber
- commentaries on idiosyncrasies in a speaker's production, stylistic and dialectal forms, etc.
- non-articulatory noises (finger-tapping, rustling of paper, etc.)
- interruptions in the recording, caused by incorrect use of the speak-button

- numbers
- abbreviations and spelling

The transcription symbols for these objects are chosen such that the objects can readily be identified and easily be distinguished from the lexical items.

### Segmentation and Labelling

For segmenting and labelling the signals, the orthographic transcription is converted into a phonological transcription representing the canonical pronunciation of the utterance. The canonical transcription is used as the basis for a list of labels. These are to be time-aligned with the signals and where necessary modified to indicate differences from the canonical form (deletions, insertions and replacements).

Segmentation is discrete and exhaustive. The segmentation of each signal file begins at the onset of utterance and ends with the cessation of utterance. The placement of a label simultaneously symbolises the beginning of one segment and the end of the previous segment.

```
g097a003.s1h
ANS003: das +/is<Z>=/+ freut mich , da"s
        Ihnen das pa"st <A> <#Klicken> .
oend
 d a s+  Q I s z: =/+  f r 'OY t  m I C+  , d a s+
 Q i: n @ n+ d a s+  p 'a s t  h:  .  :k
kend
 c:  d -h a s+  Q- -q I s z: =/+  f r 'OY t  m I C+
 , d -h a s+  Q- -q " i: n @- %n+  d -h a s+
 p -h 'a s t -h  h:  .  :k
hend
`  1809 #c:
   1809 ##d
   2994 $-h
   3379 $a
   3720 $s+
   4959 ##Q-
   4959 $-q
   4959 $I
   5662 $s
  10424 $z:
  10424 $=/+
     .    .
     .    .
```

*Figure 2: Example of a label file. From top to bottom: orthographic transcription, canonical form, the modified labels after segmentation and part of the labels with their sample numbers.*

The product of the segmentation and labelling is a text file containing the orthographic and canonical phonological transcriptions and a list of (modified) labels and their times (see Figure 2).

The system for segmenting and labelling was originally developed for read speech [5]. It has been extended to include labels and conventions for the objects introduced for spontaneous speech. As with phonetic-phonological labels, the new ones are time aligned with events in the signal [6].

In addition, a system for prosodic labelling is at present being developed [7].

## ACKNOWLEDGEMENT

This work was partially funded by the German Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01IV101M7. The responsibility for the contents of this study lies with the authors.

## REFERENCES

[1] Karger, R., Wahlster, W. (1994), *VERBMOBIL Handbuch*, Verbmobil Technisches Dokument 17, Saarbrücken: DFKI.

[2] Pätzold, M., Simpson, A. (1994), *Das Kieler Szenario zur Terminabsprache*, Verbmobil Memo 53, Kiel: IPDS.

[3] Kohler, K.J., Lex, G., Pätzold, M., Scheffers, M., Simpson, A., Thon, W. (1994), *Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0*, Verbmobil Technisches Dokument 11, Kiel: IPDS.

[4] *Der Duden*, (1991), 20th ed., Mannheim, Wien, Zürich: Dudenverlag.

[5] Kohler, K.J. (1994), *Lexica of the Kiel PHONDAT Corpus: Read Speech*, vol. I, AIPUK 27, Kiel: IPDS.

[6] Kohler, K.J., Pätzold, M., Simpson, A. (1994), *Handbuch zur Segmentation und Etikettierung von Spontansprache - 2.3*, Verbmobil Technisches Dokument 16, Kiel: IPDS.

[7] Kohler, K.J. (1995), "PROLAB - The Kiel System of Prosodic Labelling", *Proc. XIIIth ICPhS.*