An Implemented Method for Distributed Collection and Assessment of Speech Data

Alexander Siebert, David Schlangen *and* Raquel Fernández
Department of Linguistics
University of Potsdam, Germany
{das|siebert|raquel}@ling.uni-potsdam.de

Abstract

We present an approach to decreasing the cost of collecting speech data by a) distributing experimental setups as a downloadable computer program that records data and sends it back to an experiment server and b) by 're-using' subjects for instant quality evaluation of the collected data. As an example of the kind of settings in which this approach can be used, we also shortly describe an experiment we have conducted; evaluation of the collected data showed no negative effect of the 'unsupervised' collection method.

1 Introduction

While running experiments in a distributed fashion over the Internet has become accepted practice in Psychology, this methodology has so far rarely been adopted where collection of speech data is involved.[1] In the work reported here, we wanted to make available the advantages of online experimentation that are often cited (the following list is adapted from (Birnbaum, 2001)) to speech data collection:

- Freedom from the constraints of testing people at a particular time and place;
- Automatic coding and construction of data files (no data entry by assistants);
- Opportunity to obtain large and heterogeneous samples;
- Possibility to conduct cross-cultural research without the expense of travelling;
- Reduced costs of experimental assistants.

Collecting speech data poses additional technical challenges; the usual problems with data collected in this way (reliability; self-selection of subjects; data quality) also have to be addressed. The methodology we have devised (and implemented) to tackle these questions will be described in the next section. As a concrete example of an experimental setting which profits from this approach we briefly describe in Section 3 a data collection we conducted. We close with a discussion of related work (Section 4) and planned future work (Section 5).

2 Distributed Data Collection

In this section we describe the data collection methodology and the implementation we have built. We describe both in rather abstract terms here to underline the generality of the approach; a more concrete example is to follow in the next section.

2.1 Methodology

The approach is probably best explained by running through one data collection cycle. Figure 1 illustrates the data flow through the different steps. First (Step 0), the subject signs up for the experiment, using a form presented by the (web-)server. At this point, eligibility tests can be executed to filter out subjects that do not fit criteria that experimenters might want to set (e.g., first language, handed-ness, etc.).[2] Successful applicants then get access to the

[1] See e.g. (Birnbaum, 2001) for an introduction to conducting psychology experiments over the Internet, and the discussion below in Section 5 for speech-related work.

[2] A technical factor that limits the pool of potential subjects is that broadband Internet access (for down- and uploading materials) and a headset (for recording) is required on the side of

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 47–50,
Antwerp, September 2007.

experiment software. The software at this point does not contain the actual experiment script, which is only downloaded when the subject starts the actual experimental run (Step 1). The script, which controls the stimulus items, the order in which they are presented, and also the data that is to be evaluated in Part II (see below), is created on-the-fly by the server (Step 2), according to what is needed in the current state of running the experiment.

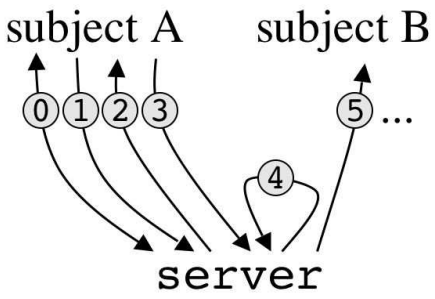

Figure 1: The Data Collection Cycle

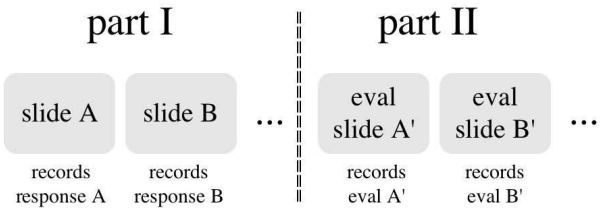

Figure 2: Schematic View of One Run

Figure 2 shows schematically one run of the experiment software for one subject. The software presents a number of "slides" to the subject and records her reactions. These "slides" can contain static information (e.g., text to read out, instructions to follow, etc.) but can also offer interactive content (e.g., puzzles to solve by manipulating items, or questionnaires); the reactions to record can range from GUI events (e.g. mouse clicks) to audio, and the responses can be timed at sub-second accuracy level. (In psychology terminology, a slide would be a single stimulus, and the recorded reaction would be the response.)

In Part II of the experiment, and this to our knowledge is an entirely novel strategy, material recorded from other subjects can be presented to the current subject, together with an evaluation questionnaire. E.g., in a simple recording experiment where the slides just contain sentences to read out, this phase II would consist of presenting to the current subject the pairs of slide and recording from a previous subject. The task then would be to evaluate the quality of the recording (or even whether the audio indeed contains a reading of the sentence!).[3]

Finishing the run brings us back to Figure 1, and Step 3, where the collected data is sent back to the experiment server. In this step audio data can optionally be compressed (lossy into MP3 format or lossless using `bz2`) to reduce the amount of data to be transferred. Step 4 then implements a consistency check. If there are criteria to do so, the data from Phase I might be pre-checked (e.g., recordings whose length deviates significantly from some preset threshold or from the mean of the data collected so far), and also the evaluation data from Phase II can be checked. The goal here is to flag all (and only) "suspicious" data, which can then be checked by the experimenter, while trying to keep as much of the data collection as possible running without further intervention.

In Step 5 finally the cycle starts again for a different subject, this time with subject A's data being available for evaluation in B's Phase II.

2.2 Implementation

On a more technical level, the data collection tool proper can be seen as a GUI shell that organises the advancement of the "slides", makes available facilities for recording data (audio, timings, GUI events, etc.), and presents data for quality assessment / evaluation. The presentation of the actual content of the slides is left to code that interfaces with this shell. (We are currently working out the best way of making this interface as general as possible; the release version will at least include an option for simple display of static content and as an example the code used in our data collection described below.)

In Phase II, the tool offers comprehensive audio controls to the user (a position slider and the usual tape-deck controls), it also allows to record all use

the user. However, in 2007 these are not unrealistic requirements.

[3] In a way we're taking our cue here from community websites that allow users to evaluate other users' contributions and hence collectively rank them.

the subject makes of these controls (see discussion of our example task below in Section 3.3).

The tool is implemented in C++ using the QT toolkit (for platform independence). It runs on Windows and Linux computers (there currently are problems with the audio library on Apple Macintosh) which must be equipped with a soundcard and headset. It weighs in at less than 5MB—a tolerable download.

3 An Example: Collecting Puzzle Moves

In this section we describe the setting for which we initially built the tool; it is at the more complex end of the spectrum of possible uses and hence nicely illustrates the potential of this strategy.

3.1 Collecting Data

The project in which this approach was developed is interested in modelling a puzzle task at both the content level, where one of the questions is how reference is made to pieces of the puzzle, and at the coordination level, where one of the questions is how different levels of interactivity shape the conversation.

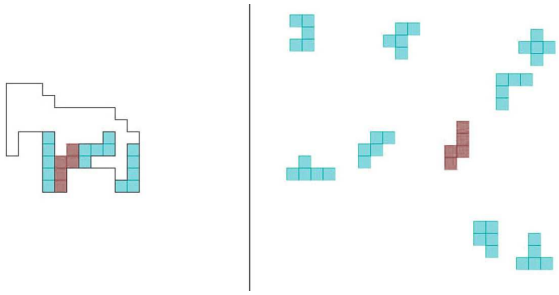

Figure 3: Example Pentomino Scene

More concretely, the task given in the data collection described here consists in describing verbally moves in a Pentomino puzzle game. Figure 3 presents one example scene; the move that is to be described here involves naming the highlighted piece on the right, describing the necessary rotation operation, and finally describing the target location in the outline on the left. This is Phase I in the terminology described above. In Phase II then scenes are presented without highlights and the recorded commands of other subjects are played, the task being to execute these commands (i.e., identify piece, rotate it, and identify target location) and then to indicate the confidence in the action performed. The audio is presented through the player tool described above and all actions (pause, repeat, skip) are recorded, as well as the judgement and the actual correctness of the execution.

Using our tool, we presented 30 scenes for execution and as many scenes for evaluation to 10 subjects (native German speakers; mostly university students). This resulted in 210 minutes of audio material, 9 sets of evaluation judgements, and a large amount of additional behavioural data (actions during evaluation).[4] The mean length of one scene description was 41 sec, with successfully followed descriptions being significantly shorter than those that couldn't be followed. Of the latter there were only 36 (12%), however, which indicates that the subjects took the recordings task seriously and produced valuable data.

As this is only a very indirect evaluation of the methodology, we also compared the audio quality of the collected recordings with that of recordings from the corpus described in (Schlangen and Fernández, 2007), which were collected with similar equipment (consumer-level headsets) but in controlled studio conditions. We used as our metric for comparison the "speech to noise ratio" as computed by the `stnr` tool from the NIST Speech Quality Assurance Package,[5] and, quite interestingly, found no significant differences between the corpora.

In the following we describe briefly two questions we addressed with these data.

3.2 Learning visual semantics

One of the goals of our project is to bridge natural language semantics, in particular for referring expressions, to perceptual features (along the lines of e.g. (Roy, 2002)). To this end, we need a large number of descriptions in our domain. The interactive material we have recorded in a different experiment (Fernández et al., 2007) provided some, but proved time-consuming to collect, annotate and segment, which is why we set out to collect more

[4] There's an obvious catch in the methodology we haven't mentioned yet: when the first subject does her run, there isn't any data available to evaluate yet. In our case, we separated for the first subject phase I (collection) and phase II (assessment).

[5] Available from `http://www.nist.gov/speech/tools/index.htm`.

in a non-interactive setting. The quality assessment data reported above convinced us that the descriptions collected in this way were not worse than those collected in the interactive setting.

Using a simple set of visual features and a simple vector-based learning and recognition model implemented as a baseline (aligning nouns with vectors of visual features; class / reference of test items determined by minimal distance) already achieved an accuracy of 62%.[6]

3.3 'Interactivity' in a non-interactive setting

In (Fernández et al., 2007) we ran the puzzle experiment in a fully interactive setting and in one with restricted interactivity (push-to-talk). The completely non-interactive material collected here gives us a good further comparison. We were especially interested in the use subjects made of the player tool to recreate some semblance of 'interactivity' through stopping, skipping and repeating audio material. The analysis of this is still going on.

4 Related Work

As mentioned in the introduction, conducting experiments over the Internet is common practice in Psychology these days (Birnbaum, 2001; Reips, 2002),[7] However, these experiments rarely involve audio. (Font Llitjos and Black, 2002; Black and Tokuda, 2005) present experiments on collecting *evaluations* of speech over the Internet; SpeechRecorder (Draxler, 2006) offers recording over the Internet much like our system, but with no provisions for recording other behavioural measures like reaction times. The combination of experiment / collection with instant user-based quality assessment that our approach offers is, to our knowledge, novel.

5 Conclusions and Future Work

We have presented an implemented methodology for distributed collection of speech data. The implemented tool is flexible in the kind of stimuli that can be presented (static and dynamic) and can record audio and other behavioural data (with sub-second accuracy). As a novel strategy for overcoming reliability problems connected to "unsupervised" data collections it allows for immediate, equally "unsupervised" quality assessment. We believe that there is a wide range of use cases in which the tool can support collection of spoken data, e.g. recording "think aloud" protocols for cognitive tasks, collecting domain utterances with simulated dialogue systems, and many more.

We are currently exploring ways of letting the software run in the user's web-browser (using Flash, or AJAX-style programming) rather than as an independent executable, but first experiments indicate that this cannot yet provide the timing accuracy and reliability that our current tool has reached.[8]

References

Michael H. Birnbaum. 2001. *Introduction to Behavioral Research on the Internet*. Prentice-Hall, NJ, USA.

Alan W. Black and Keiichi Tokuda. 2005. The blizzard challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proceedings of Interspeech2005*, Lisbon, Portugal, September.

Christoph Draxler. 2006. Web-based speech data collection and annotation. In *Proceedings of 'Speech and Computer (SPECOM2006)'*, St. Petersburg, Russia, June.

Raquel Fernández, David Schlangen, and Tatjana Lucht. 2007. Push-to-talk ain't always bad! comparing different interactivity settings in task-oriented dialogue. In *Proceeding of DECALOG (SemDial'07)*, Trento, Italy, June.

Ariadna Font Llitjos and Alan Black. 2002. Evaluation and collection of proper name pronunciations online. In *Proceedings of LREC2002*, Las Palmas, Canary Islands.

Ulf-Dietrich Reips. 2002. Standards for internet-based experimenting. *Experimental Psychology*, 49(4):243–256.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3).

David Schlangen and Raquel Fernández. 2007. Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech 2007*, Antwerp, Belgium, August.

[6] More detailed results will hopefully soon be reported.

[7] See also `http://psych.hanover.edu/ ... research/exponnet.html` for an up-to-date list of open experiments.

[8] **Acknowledgements:** The work reported here has partially been funded by EU (Marie Curie Programme) and DFG (Emmy Noether Programm). Thanks to the anonymous reviewers for their helpful comments.