The FASiL Speech and Multimodal Corpora

Hans Dolfing[1], David Reitter[†2], Luis Almeida[3], Nuno Beires[3], Michael Cody[2],
Rui Gomes[3], Kerry Robinson[1], Roman Zielinski[4]

[1]Vox Generation Ltd., London, UK
[2]MIT Media Lab Europe, Dublin, Ireland
[3]Portugal Telecom Inovaçao, Porto, Portugal
[4]Capgemini Sweden, Stockholm, Sweden

`hdolfing@voxgen.com`

Abstract

In the context of the FASiL project, we have studied natural language interactions in a unimodal (speech only) and multimodal (speech and graphics) interface to a personal information management database. We collected multilingual corpora to investigate these interactions in Portuguese, English and Swedish. The corpora are used to train language models, to update acoustic models, to study semantic concepts, multimodal interactions, and dialogue management strategies. The corpora are annotated in a uniform way, with timings, transcriptions, and semantics. We report on the structure and design of the corpora which are now available via ELRA.

1. Introduction

In the FASiL project, we investigated unimodal and multimodal interactions based on the example application of a virtual personal assistant (VPA) that helped users to manage their emails, calendar and address book. Within this framework, we developed systems that interact with users in Portuguese, Swedish, and English to evaluate and demonstrate techniques of multimodal fission and fusion [1], language modelling, semantics [2, 3], and dialogue modelling [4].

The FASiL corpora collected during the project life contain recorded interactions with the VPA to investigate the above topics and are made available as five datasets: one speech corpus for each language, a combined speech corpus containing all three languages, and a multimodal corpus.

The data in different languages is coordinated in terms of content. Native speakers were asked to perform the same high-level tasks, such as to schedule a meeting. Similarly, the basic capabilities of the systems in each language were the same. The coordination makes the corpus a valuable dataset, not only for typical speech recogniser improvement but also for comparative studies of linguistic phenomena. The FASiL speech corpora consist mainly of conversational, spontaneous speech as opposed to the data from SpeechDat [5]. With respect to the VerbMobil [6] corpora, the main differences with the FASiL corpora are in applications, modalities, languages and recording channels. The FASiL speech corpora have already been used in the FASiL project for acoustic adaptation of the involved recognisers, for language model training, for quality assurance on yes/no utterances, and dialogue design.

The multimodal corpus contains simultaneous and sometimes co-ordinated interactions via a graphical and voice-driven user interface. This corpus was used to develop a multimodal demonstrator and investigate and train models for the fusion of multimodal interaction events in a VPA context. The corpus is a continuous source of data for recogniser improvement, and study of linguistics and multimodal phenomena such as fusion of GUI and VUI events. The availability of parallel language versions may represent an advantage over other corpora, such as QuickSet [7]. In comparison to the German-language corpora from SmartKom [8], FASiL has a different area of application and a slightly different choice of modalities.

In the rest of this paper, we report on the structure and content of the unimodal (speech-only) and multimodal corpora. These corpora are now available via ELRA[1] to build multilingual and multimodal applications, study semantics in a multilingual environment, and build acoustic and language model data for speech recogniser improvement. In the following Sections 2 to 6, we discuss the methodology, recording, storage format, and content of the corpora.

2. Methodology

Lacking a ready-to-use speech-driven computer system to collect the data with, speech and multimodal interaction modes were simulated using Wizard-of-Oz (WOz) experiments of differing levels of complexity.

For the collection of the unimodal speech corpus, subjects were asked to complete a number of email, calendar and address-book tasks by giving instructions to a human operator over the phone. The operator attempted to carry out the subjects' instructions using a personal information management (PIM) database that was populated with fictitious emails, appointments and contacts. They were provided with a list of functions that they could perform, which prevented them from being too proactive in order to mimic the functional capabilities of a computer system. Operators, however, were not given the precise wording of the allowed responses. While the subject could not actually see the operator, they could hear their natural voice.

For the collection of the multimodal corpus, subjects interacted with what appeared to be a fully functioning computer system. In contrast to the unimodal study, users heard synthesized speech, and viewed graphical responses on a screen. Behind the curtain, however, all system responses were generated

[†]Now at School of Informatics, University of Edinburgh, UK

[1]European Language Resource Association, http://www.elra.info/

in real time by human operators (wizards) using the same fictitious PIM data as for the unimodal collection.

The constraints in these experiments were meant to elicit user input that is representative of what is to be expected in an actual system. While it is well-known that users may show an emotional response to a computer just like they do to a human [9], WOz allows us to justify a simulated lack of higher-level intelligence.

2.1. Experimental Conditions

Three sites in Ireland, Sweden, and Portugal conducted the WOz studies using 100 subjects recruited locally. Subjects were professionals and students (average age: 34.6, standard deviation: 10.6), had prior experience with PIM software such as *Outlook*, but no specific training in spoken user interfaces. Subjects completed a fixed list of tasks such as "arrange a meeting with Veronica for tomorrow at three to discuss the new recruitment procedures". Of the 100 local subjects, 70 subjects concentrated on the unimodal WOz study and 30 subjects completed tasks in a unimodal (voice only) and multimodal condition and in the presence/absence of a noise distractor. Wizards received training; training materials as well as subject tasks were standardized across experimental sites.

Anecdotal evidence from a questionnaire completed by participants after the multimodal experiments suggests that they found the system sufficiently easy to use, and that the majority of them were not aware that the interaction was actually based on a simulation, rather than a fully functional system. This is somewhat surprising as the system offered near-perfect speech recognition and understanding, which is beyond the abilities of any state-of-the art system.

3. The WOzOS system

In many cases, WOz systems have been purpose-built to investigate specific system designs [10, 11, 12]. More generic platforms allow simulation-based research in a wider variety of contexts. In NEIMO [13], Apple's *HyperCard* was used to assemble an UI in real time. The system allowed both GUI elements and video, but not sound, to be transmitted and logged. In SUEDE [14], a speech interface is generated from a relatively simple toolkit.

In the unimodal WOz phase, the wizard manipulated Outlook and the user interacted only over a voice channel. Several unimodal platforms were studied and deployed at each site, complying with recording requirements for the corpus collection, i.e, preferably independent recording of the conversation channels and directly from the telephone line.

After our experience with the unimodal WOz system in three languages, the Wizard-of-Oz Operating System (WOzOS) was implemented to facilitate the multimodal WOz. It is a Java-based distributed platform to support simulation-based research across a variety of situations. It combines standard GUI elements with spoken input and synthesized voice output. System output is assembled in real-time by two wizards.[2]

WOzOS runs on three networked workstations: Operator, Session Manager (both wizards) and Client (subject). Usually, the client system will be physically separated from the wizards' workstations, as done during the FASiL corpus collection.

FASiL used a commercial text-to-speech (TTS) module (ScanSoft's *RealSpeak*) to synthesize spoken output, but the use of a high-quality free TTS module (*Festival*) is possible. We used a head-mounted microphone to capture the subject's input and headphones to provide speech output to the subjects. The data for the scenario to be simulated comes from widespread PIM software: Microsoft Outlook. Its database contained a prepared collection of emails, contacts and appointments, which was reset before each experiment.

A credible system must respond quickly; excessive delays could influence subjects' impression of whether they are interacting with a computer system [10]. Therefore the *Session Manager* and *Operator* share the workload in preparing output: the **Operator** retrieves data from an application such as *Outlook*, and prepares the visual representation. In parallel, the **Session Manager** assembles the speech output and ensures that all experimental tasks are carried out.

Through this methodology, we achieved our initial target of a response time of a few seconds: average response time[3] measured in a sample session was 5.5 seconds, with standard deviation 4.7.

4. Corpus structure

Each subject was recruited to participate in the WOz experiment which had two phases. The first phase, with 70 users in each language, was unimodal (speech only) and relatively unstructured. The second phase with 30 users in each language introduced multimodality and was more structured than the first phase. The multimodal interactions with speech and pen gestures were recorded by the WOz platform. Both unimodal and multimodal recordings include all event timings such as the start and end of speech input, prompts, and pen clicks. All recordings were transcribed.

While the unimodal WOz experiments resulted in a total of four corpora (a combined corpus and three language-specific ones with each about 15000-20000 utterances) the multimodal WOz experiment resulted in the corpus "fasil-mm", containing more than 90 multimodal interactions in the three project languages with a total of 15GB of data (see Table 1). The unimodal corpus "fasil-all" contains recordings of 210 users in all three languages.

Table 1: Corpora names, descriptions, and sampling details. "fasil-mm" is the multimodal corpus. "fasil-all" is the combination of the three speech-only, language-specific corpora.

Name	Nr. of users	Male/Female	Nr. of hours	Recording
fasil-pt	70	39/31	≈ 20	8kHz/8bit
fasil-sv	70	54/16	≈ 27	8kHz/8bit
fasil-uk	70	49/21	≈ 17	11kHz/16bit
fasil-all	210	142/68	≈ 64	8/11kHz
fasil-mm	92	62/30	≈ 42	16kHz/16bit

5. Recording procedure

For the unimodal corpus, we recorded as much as possible via real telephone connections, using separate channels for wizard and user where possible. The multimodal system used desktop microphones, a recording system as in Figure 1, and the WOzOS software platform discussed in Section 3.

---

[2]WOzOS has been developed in FASiL by MIT Media Lab Europe and is available free of charge as open source distribution at http://wozos.sourceforge.net/

[3]Response time being defined as time difference between end of Client input and start of a system (Wizard generated) response. This includes occasional filler phrases such as *OK, let's see...*
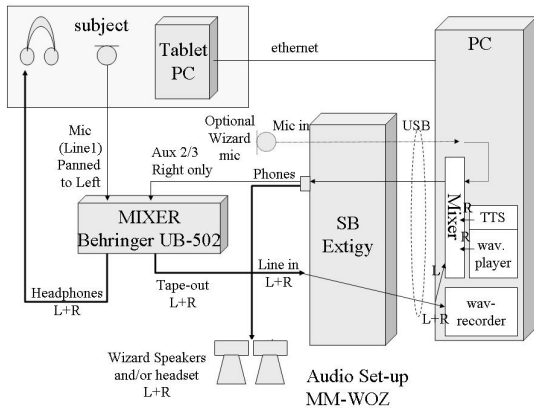
Figure 1: Swedish multimodal WOz recording setup. English and Portuguese setups differed slightly not affecting the outcome.

To avoid the risk of losing parts of utterances, the whole session was recorded and then segmented in a per-utterances basis. Utterances are identified as turns in dialogues. In some cases when the user changes the topic, the two separated utterances may be identified without a turn to the other party. Precaution was taken to avoid cutting the file in the middle of an utterance.

As with all distributed systems, the reliability of time-stamps should be considered limited; those in the multimodal corpus, however, were post-processed in order to account for offsets in timing mechanisms and small network delays.

6. Data representation

6.1. Data formats

In order to represent all events and data plus timings in the corpus, we represent each recorded session as a 'lattice'. Each edge in the lattice is an event, with start and end time, and represents a user utterance, a TTS prompt, pen click or similar. Each edge is optionally annotated with additional attributes. We represent full unimodal and multimodal sessions with a 'masterfile', loosely following the HTK Standard Lattice Format (SLF).

Table 2: Example unimodal masterfile.

S=0	E=0		T=012_0000.txt	A=header
S=10.1	E=20.0	V=012_0001.wav	T=012_0001.txt	A=prompt
S=30.2	E=35.4	V=012_0002.wav	T=012_0002.txt	A=answer

In Table 2, we see that a masterfile, with session id 012, starts with a header file, which contains the person and session specific data. The header is annotated with attribute "A=header" and has zero length in time. Next, we have the opening prompt of the system, recorded in sound file "V=012_0001.wav", and annotated with "A=prompt" as system prompt. The start time "S=10.1" and end time "E=20.0" indicate a long prompt of about 10 seconds. The prompt is followed by a user utterance annotated with "A=answer" and transcription "T=012_0002.txt".

We have complete timings of the WOz sessions and we also know what are prompts or user utterances. All data files of a session "N" are stored in directory "N".

The multimodal corpus uses masterfiles such as in Table 3. In addition to the voice interaction, we provide screen output, stored in picture form, and touchscreen stylus clicks as input. In Table 3, the 'click' event is overlapping in time with the voice input. The event at time 60.1 is starting screen output, where the corresponding text or instruction is stored in the text file indicated with "O=".

The multimodal dataset contains additional annotations that describe referring expressions and high-level discourse semantics and is delivered in the XML-based TASX format [15].

Table 3: Example multimodal masterfile.

S=0	E=0		T=012_0000.txt		A=header
S=10.1	E=20.0	V=012_0001.wav	T=012_0001.txt		A=prompt
S=30.2	E=35.4	V=012_0002.wav	T=012_0002.txt		A=answer
S=31.7	E=32.4	A="user input: mouse: Click 132,97"			
S=60.1	E=62.2	GR=012_0004.gif		O=012_0004.txt	A=screen-output

6.2. Transcription and Annotation

We defined a word-for-word transcription for dialogue between subject and wizard that covers adequately the most important acoustic events for the training and testing of the VPA's language models [3]. The transcription is orthographic and the utterances are represented as lower case without punctuation. It includes marks, in uppercase, that represent audible acoustic events (non-speech and irrelevant speech) present in the corresponding waveform files.

The character set used for the orthographic transcriptions is ISO-8859-1. More complex schemes exist for representing this information in more detail, but their accurate use requires significant training, expertise and time if, but does not necessarily support purposes in applications similar to the FASiL VPA.

6.2.1. Unimodal transcription

The transcription represents the actually spoken words as lower case without punctuation (no question marks, commas, etc.), using the standard spelling conventions of each language. If there are words that can be spelled in two or more different forms, one of these forms was established as the standard for annotation purposes. Contractions like "I'm" and "I'll" remain.

The transcription of spelled words and abbreviations is orthographic not phonetic. They are represented as letter sequences in lower case separated by a space. For example the transcription of "FCP" should be: "f c p". But, if the speaker says: "FC Porto" the transcription is "f c porto".

Number sequences such as times, dates or money amounts must be transcribed just as they were said, e.g., "two thirty p m" or "five o'clock". The same rule is applicable to email and web addresses. For instance, "www dot euronews dot com".

Spoken forms with wrong pronunciations, missing endings, and similar are represented by their correct orthographic form, as both language model text training and acoustic training in FASiL ignore incorrect words. Intelligible word fragments, i.e., instances in which the speaker did not complete a word, were considered mispronunciations. Words or speech segments that are completely unintelligible are transcribed as "NOISE".

The Portuguese and Swedish data contain a few English words like "forward". These words were orthographically transcribed in English. The mix of Swedish, Portuguese and English words will be part of any resulting language model.

Non-speech acoustic events are transcribed as follows: "NOISE" for parts of words, coughing, telephone noise and the like; "UH" for any hesitation, e.g., "um", "err", "mmm" and equivalents. Text irrelevant for the domain, e.g., message bodies or comments about the experiment, is represented as "IRRELEVANT". Pauses were not marked explicitly as it is possible to extract the pause lengths automatically when necessary.

6.2.2. *Multimodal transcription*

An important difference is the annotation of the unimodal and multimodal corpora. In addition to the annotation rules described above for the unimodal corpus, the multimodal corpus also includes semantic annotations. As a tool, we used the TASX [15] platform to visualize the sound files and annotations. The semantic annotations focus on dialogue goals and referring expressions.

We distinguish two tiers of dialogue acts: 9 high-level dialogue goals represent a form of interaction with the data in the database, e.g., *Find a contact*, *delete email*, *create appointment*. We devised 11 types of more immediate interaction with the user interface that mark up the input of dates, times, names, numbers, email addresses and the like.

Referring annotations are heavily used in naturally occurring language, and they play an important role in multimodal human-computer interaction. Examples include deictic expressions such as *these people*, which are accompanied by pointing gestures on the touch-screen, or anaphoric pronouns such as *it*, referring to an element in the (previous) dialogue context. Deixis can be combined with a verbal specification, as in the example of a *deictic definite* mentioned before, or it can stand alone, as in the deictic pronouns *this* and *that*. We annotate all referring expressions, including the fully-specific forms, such as in names. Unique identifiers are used to annotate the binding relationships between different referring expressions, that is, they identify the discourse entity that the WOz system or subject has referred to.

An example transcription would be "Send [Pronoun 7 it] to [Full 88 Liz Dixon]", where "it" is a pronoun referring to an email that has been mentioned before and is assigned the binding index 7, and "Liz Dixon" is a person entry with id 88 in the database. The binding indices allow for training and evaluation of multimodal fusion and anaphora resolution algorithms. We provide such annotations in English and Portuguese.

To evaluate whether the annotation task was well-defined and annotators were well-trained to produce consistent results, we had an additional annotator redundantly produce data for two sessions in English and calculated Kappa [16] as a measure of inter-annotator agreement level. Very good agreement ($\kappa > 0.8$) was reached for dialogue goals, interaction and referring expression types.[4]

Further work on these data could include annotations for the information displayed on the screen. Currently, we provide screenshots of the interaction situation, but no structured information of particular discourse entities. What is shown on the screen likely influences the way people refer to the items. Corpus-based studies in the HCI context become easier with structured annotations of what is visible.

7. Conclusion

We presented the speech and multimodal corpora developed in the FASiL project. They have already been successfully used to improve and adapt acoustic models, to develop new types of language models, to prime research on adaptive dialogue management and to investigate models for multimodal fusion and linguistic reference, for which transcriptions and selected semantic annotations are provided. The corpora are available via ELRA.

---

[4]All of the multimodal annotations were carried out in Dublin. Principle annotators had native or near-native command of the respective languages and received prior training.

9. References

[1] Reitter, D., Panttaja, E., Cummins, F., "UI on the fly: Generating a multimodal user interface", Proc. Companion HLT/NAACL, Boston, USA, 2004.

[2] Buckley, P, Horowitz, D., Lal, P., "A maximum entropy shallow functional parser for spoken language understanding", ICSLP, 2149-2152, Jeju, South Korea, 2004.

[3] Dolfing, H., Buckley, P., and Horowitz, D., "Unified language modeling using finite-state transducers", ICSLP, 1049-1052, Jeju, South Korea, 2004.

[4] Robinson, D., Horowitz, D., Bobadilla, E., Lascelles, M, Suarez, A., "Conversational dialogue management in the FASiL project", Proc. 5th SIGdial Workshop on Discourse and Dialogue, 19-22, Boston, USA, 2004.

[5] Höge, H. et al, "SpeechDat multilingual speech databases for teleservices: across the finish line", EuroSpeech, Budapest, Hungary, 1999.

[6] Wahlster, W. (ed.), "Verbmobil: Foundations of speech-to-speech translation", Springer.

[7] Cohen, P. et al, "Quickset: multimodal interaction for distributed applications", Proc. ACM Multimedia, 31-40, Seattle, USA, 1997.

[8] Wahlster, W., "SmartKom: Multimodal communication with a life-like character", EuroSpeech, 1547-1550, Aalborg, Denmark, 2001.

[9] Reeves, B. and Nass, C., "The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places", University of Chicago Press, Chicago, 1996.

[10] Oviatt, S., et al., "A Rapid Semi-Automatic Simulation Technique for Investigating Interactive Speech and Handwriting", ICSLP, 1351-1354, Banff, Canada, 1992.

[11] McInnes, F.R., et al., "User responses to prompt wording styles in a banking service with a Wizard of Oz simulation of word-spotting", Proc. of IEE Colloquium on Advances in Interactive Voice Technologies for Telecommunications Services, 1-6, London, UK, 1997.

[12] Wyard, P., Churcher, G., "A Realistic Wizard of Oz Simulation of a Multimodal Language System", ICSLP, 1351-1354, Sydney, Australia, 1998.

[13] Coutaz, J., Salber, D., Carraux, E., "NEIMO, a Multimodal Usability Lab for Observing and Analyzing Multimodal Interaction" Proc. Companion CHI'96, Vancouver, Canada, 1996.

[14] Klemmer, S., Sinha, A., Chen, J., Landay, J., Aboobaker, N., Wang, A., "SUEDE: A Wizard of Oz Prototyping Tool for Speech User Interfaces", CHI Letters: Proc. ACM Symposium on User Interface Software and Technology, 1-10, 2000.

[15] Milde, J.T., Gut, U. "The TASX-environment: an XML-based toolset for time-aligned speech corpora", LREC 2002, 1922-1927, Las Palmas, Spain, 2002.

[16] Cohen J., et al.,"A coefficient of agreement for nominal scales.", Educational and Psychological Measurement, 20, 37-46, 1960.