VERBMOBIL: THE COMBINATION OF DEEP AND SHALLOW PROCESSING FOR SPONTANEOUS SPEECH TRANSLATION

Thomas Bub[*], *Wolfgang Wahlster*[*], *Alex Waibel*[+]

[*]German Research Center for Artificial Intelligence GmbH (DFKI), Stuhlsatzenhausweg 3 (Bau 43),
D-66123 Saarbruecken

[+]School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213

bub@dfki.uni-kl.de, wahlster@dfki.uni-sb.de, ahw@cs.cmu.edu

ABSTRACT

Verbmobil[1] is a speech-to-speech translation system for spontaneously spoken negotiation dialogs. The actual system translates 74.2% of spontaneously spoken German input.

In the following we give an overview of the *Verbmobil* system. After the introduction of the *Verbmobil* scenario and the unique constraints of the project, we describe the underlying system architecture and its realization. The progress that was achieved on the end-to-end translation rate owes much to the increase of the word recognition rate from 45% in 1993 to 87% in 1996. But in order to achieve the envisaged coverage on the incertain speech recognizer output, deep and shallow approaches to the analysis and transfer problem had to be combined.

1. INTRODUCTION

Verbmobil is a speech-to-speech translation system for face-to-face negotiation dialogs [1]. The Verbmobil scenario assumes a native speaker of German and a native speaker of Japanese. Both possess at least a passive knowledge of English. Thus the conversation may proceed in English. In case the active knowledge of English turns out to be insufficient, the dialog partners are allowed to use their native language. The *Verbmobil* system supports them by translating from their mother tongue, i.e. Japanese or German, into English.

The project comprises expertise from the domains of signal processing, computer linguistics and artificial intelligence. It includes 29 partners providing 150 researchers and engineers. The software has been developed at different sites in Germany, Japan and in the United States and has been integrated by a central group of highly skilled software engineers.

The continuous system and module evaluations document, that during the past four years significant progress has been achieved. • The acoustic word error rate decreased from 55% to 13%.

• The actual speech-to-speech translation rate is 74.2% in the domain of appointment scheduling dialogs.

• The actual wordlist contains 2461 words compared to 610 in the first integrated system.

• System performance was reduced from about 20 times signal length to 5.7 times signal length.

2. THE ARCHITECTURE OF THE *VERBMOBIL* RESEARCH PROTOTYPE

The rapid progress that Verbmobil achieved during the past four years owes very much to the early availability of a clearly specified, but flexible and extendible architecture.

The fundamental concept of the functional architecture is simple but efficient: The system functionality is broken down into functionally closed tasks which can be achieved by independent modules communicating with each other. From the architecture point of view the system thus consists of a number of functional black boxes, such as *speech recognizer*, *syntactic-semantic analysis*, *transfer* or *generation*, that exchange information. These functional modules are coordinated and controlled by a handful of technical modules.

The concept of communicating modules provided solutions for most of the problems the project had to overcome:

• Existing software could be adapted and reused, thus permitting the partners to build upon their existing know-how and software.

• The geographically distributed software development became possible because the system could be divided into encapsulated functional modules.

• The number of interface formats between the functional modules could be reduced to only four different data types, allowing for the autonomous development and testing.

• Based upon these concepts the system functionality evolved by iteratively exchanging single modules or groups of functional modules (*breadboard architecture*).

[1] This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 413-4001-01 IV 301 4. The responsibility for the contents of this paper lies with the authors

- Alternative realizations of modules or even of groups of modules could be tested in the context of the complete operational system.

2.1. The Functional Architecture

The functional architecture of *Verbmobil* is a multiagent architecture. Autonomous modules process speech and language data by interacting with other modules, if additional information is needed. The functional architecture of the *Verbmobil* research prototype consists of 43 functional modules (figure 1).

According to the above mentioned scenario *Verbmobil* operates in a passive *listening* mode and in an active *interpreting* mode. While the dialog partners speak in English the *keyword spotter* module (or alternatively a continuous E*nglish speech recognizer*) looks for specific keywords that are used by the dialog processing module in order to assign the relevant speech acts. Once a partner presses the *Verbmobil* activation button, the system turns into active mode and the input is processed by a set of acoustic modules including prosodic and morphologic analysis and by different concurrent *linguistic analysis and transfers*. Finally the *selection* module decides on the most suited output, depending on the actual requirements to the system (e.g. processing speed vs. quality of translation).

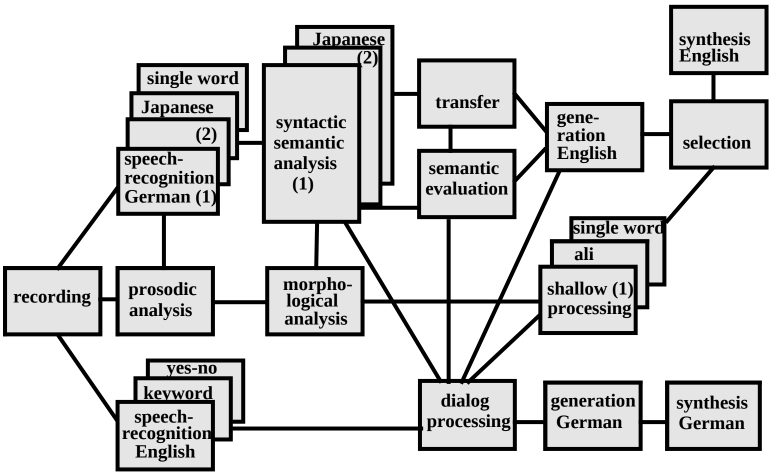

Figure 1: The functional architecture consists of communicating modules.

Acoustic Analysis. As long as the user holds the *speech button* pressed, the speech signal is recorded, digitized (16 kHz, 16 kbit) and incrementally sent to the respective Japanese or German *speech recognizer*. The module *recording* drives and controls the external recording devices (microphone, A/D-converter) and processes signal failures.

Based on the digitized signal the speech recognizers produce scored hypotheses stating which word might have been spoken in a certain time frame. These are exchanged with the prosodic and morphologic analysis as *word lattices*, the density and the size of which is configurable. Most of the recognizers, namely the German, Japanese, English *continuous speech recognizers* and the German *single word recognizer*, are based on statistical *language models* and *lexicons* containing orthographic and phonetic information. The actual size of the German lexicon is 2.461 words. Two functionally identical German speech recognizer modules are integrated. They may be activated alternatively and can be exchanged at runtime.

The *prosodic analysis* annotates the lattices with prosodic information i.e. with accents, phrase boundaries and sentence mood. This information is used by the linguistic modules of the system for the segmentation of turns and for disambiguation. It results in an important search space reduction for the syntactic analysis and therefore significantly speeds up the system.

Deep Processing. The functional architecture of *Verbmobil* permits concurrent linguistic analyses varying in processing depth, processing speed and quality of output.

The so-called *deep processing* consists of the modules *syntactic-semantic analysis*, *transfer* and *Englisch generation* and requests additional information from the modules *semantic evaluation* and *dialog processing*. *Syntactic-semantic analysis* takes the word lattice as input, searches for syntactically correct chains and their possible syntactic and semantic readings. Information on lexical semantics can be obtained from the module *semantic evaluation*. The interleaved syntactic-semantic analysis is based on unification grammars and compositional semantics representation and results in an underspecified DRS represented as 10ary *vit*-term (for ***v**erbmobil* **i**nterface **t**erm) [2]. Currently, there are three functionally similar syntactic-semantic analysis modules in the system: two alternative analyses of German and one analysis of Japanese.

The mapping of semantic structures of the source language, i.e. Japanese or German, to the semantic structures of the target language is the task of the module *transfer*. On the basis of the transferred DRS, the *English generator* produces syntactically correct utterances in English using reversible HPSG-grammars. The translations are passed as prosodically annotated strings to the English synthesis to produce the corresponding speech signal.

The module *semantic evaluation* receives the same input as *transfer*. It currently serves three purposes: it maintains a *domain model* of relevant world objects and their dependencies, computes the current dialog act and tests for inconsistent dates such as February, 30th. There are various interfaces from *semantic evaluation* to other modules that need either dialog history information or information from the *domain model*.

In close cooperation with *semantic evaluation* the module *dialog processing* maintains a dialog history based on dialog acts. *Dialog processing* puts together the results of the passive keyword spotting mode and the active deep analysis and transfer mode. Using a statistical model of dialog act transitions, the *dialog processing* also predicts the next dialog acts. The module is also responsible for the initiation of clarification dialogs with the users, e.g. in order to inform them about inconsistent dates and to resolve critical situations [3].

In cases of system faults and clarification dialogs, the *German generator* produces prosodically annotated German text that will be synthesized by the *German synthesis*.

Shallow Processing. The linguistic processing is completed by two parallel *shallow approaches*: a speech act based translation and a schematic translation.

Verbmobil maintains a stochastic context model of the dialog based on corpus training. In cases where the acoustic analysis is incomplete or extremely uncertain, this model is used by the module *shallow processing* to produce speech act based *reductionist* translations in the form of templates filled with recognized dates.

The schematic translation is realized by means of a database containing about 20.000 entries. The module *ali* abstracts from dates and names, searches for the most similar entry in the database and outputs the relevant translation by actively translating dates and names.

Both, *shallow processing* and *ali* take word lattices as input and produce text or annotated text, i.e. they have exactly the same interfaces as the deep analysis.

The shallow and the deep analyses work in parallel, thus several translations of the same turn are available. The translations are scored by the producing modules, using stochastic information and heuristics.

The final decision which translation is best suited in the current context is the task of the module *selection*. *Selection* provides different rule sets which allow for different selection criteria, e.g. system performance versus translation quality. Moreover, if there are gaps in the deep analysis of an utterance they may be filled with shallow translations.

It is remarkable that the *Verbmobil* system does not need a central functional control module. The system consists of independent non-hierarchic autonomous speech and language processing agents.

2.2. The Software-Architecture

The functional architecture directly maps to the software architecture. Functional modules are reflected by components, which in turn are realized by at least one process which may recursively start and control further processes. To meet the software-technical requirements *additional technical components* were introduced, taking over responsibility for software control, the user interface, visualization of module interfaces and automatic testing. Technical components and functional components are identically realized as interactively communicating processes.

Verbmobil uses the *breadboard* concept, i.e. the system consisted at the beginning of dummy modules which simulated the mapping of input to the corresponding output data. This allowed for the simulation of the data flow of the complete system. Hence, system integration or system evolution was nothing else but successive replacement of modules by modules with increased functionality. A further advantage of this concept is that, once the integration framework existed, an operational system was available and could be used for the evaluation and testing of single modules in the context of the integrated system.

The Verbmobil components are implemented in 8 different programming languages. Therefore the communication between the components had to abstract from the used programming languages. This was achieved by a *layered communication concept*, ranging from low level communication protocols to an easy-to-handle process communication environment that is available from within all used programming languages.

The lowest layer is realized by PVM (Parallel Virtual Machine) [4]. It is used by ICE (Intarc Communication Environment) [5] to implement the second layer which offers the needed subset of functionality in all used programming languages. The third communication layer is programming language dependent. It reduces the programming effort for component implementers to the handling of received messages.

Verbmobil is a pure software project, the complete system runs on comercially available workstations. Thus the overall concept of communicating modules (functional architecture), communicating components (software architecture) and communicating processes (realization) provides the basis for the *reuse of both functional modules and the integration framework* in different applications.

3. THE SYSTEM EVALUATION

The *Verbmobil* breadboard architecture allows the exchange of integrated modules against new module versions. These local changes lead to the evolution of the *Verbmobil* System. Continuous evaluations have shown the growing functionality of modules, module groups and finally the end-to-end performance of the complete integrated system.

3.1. The Tools for Module and System Tests

The Verbmobil integration framework permits the simulation of module interfaces. Whenever two modules may exchange information during runtime, this information flow may also be simulated in a test environment. Moreover, the information flow between modules can be corrected or edited in order to simulate module or system behaviour in certain situations. This feature is mainly used for test or presentation purposes.

The *automatic test module* has been developed to automize the simulation of system interfaces and module interfaces. It takes a specified set of test data, e.g. the speech signals of 1000 turns, and feeds them successively into a given interface channel. Intermediate results and the final results are archived for further evaluation.

3.2. The System Evaluation Process

While we use public test data for testing during development, the regular evaluations operate on unknown data.

During the end-to-end evaluation of an integrated system version a large set of original audio data, i.e. recorded spontaneously spoken turns, is automaticly sent to the speech recognizer as system front end. Verbmobil processes the turns and the test environment archives the produced translations.

In the next step the transliteration of the audio data together with the relevant translations are presented to professional interpreters who judge the quality of the results. As Verbmobil is a geographically distributed project this is done by means of the internet, i.e. by using interactive Web pages. The final evaluation of the Verbmobil research prototype was based on the evaluation of about 20.000 turns.

In order to process the huge amounts of test results, the interpreters work on simplified evaluation criteria. A translation is considered to be approximatively correct if

- the original information content has been preserved and
- the resulting translation is understandable.

Selected modules or groups of modules are evaluated by further analyzing the archived intermediate input-output data. Whereever possible this process is either automated or supported by software tools. For example, the speech recognizer output is automatically compared to the transcription of the audio signal.

For other interfaces, e.g. the syntactic-semantic analysis, manual judgement is necessary, which is supported by suitable visualization tools.

3.3. The Evaluation Results

The statistics of the final system evaluation (research prototype version 1.0) is shown in figure 2. The evaluation was based on a randomly choosen test set. The only constraint was that the utterances should only contain words of the Verbmobil wordlist (2461 words).

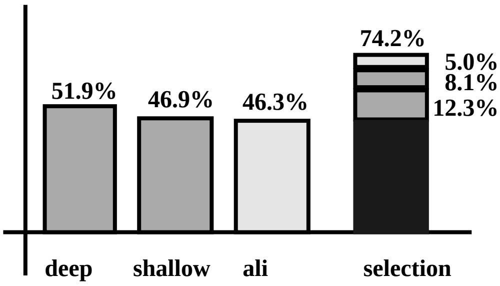

Figure 2: The end-to-end performance of Verbmobil is based on the combination of deep and shallow processing.

Whereas the overall translation rate of the system is 74.2%, the different deep and shallow approaches achieve only results between 46.3% and 51.9%. In 48.8% of the test cases at least two analyses produced a correct translation, but each approach contributes individually to the overall result. In 12.3% of the test cases only the *deep analysis* produces a correct translation, in 8.1% only the *shallow processing*, in 5.0% only the *schematic translation*.

4. CONCLUSION

The *Verbmobil* project has shown that large and complex projects can be successfully conducted even in a geographically distributed environment [6]. The underlying concepts (breadboard architecture, multiagent architecture) have proven extremely suitable. Being a pure software project, *Verbmobil* has produced reusable functional speech and language processing modules and moreover a reusable integration framework for complex heterogeneous systems, e.g. used for speech and language systems at Daimler-Benz AG and IBM Informationssysteme GmbH.

The *Verbmobil* translations have been evaluated by independent interpreters, resulting in an overall translation rate of 74.2%. This result could only be achieved due to the combination of shallow and deep analyses.

The project has achieved significant progress in the last 4 years. As one of the worlds‘ most innovative systems it combines innovative features, such as clarification dialogs, concurrency and sophisticated dialog models.

5. REFERENCES

1. Wahlster, W.: “Verbmobil: Translation of Face-to-Face Dialogs“. Proceedings of MT Summit IV, Kobe, Japan, July 1993.

2. Dorna, M.: “The ADT-Package for the Verbmobil Interface Term“. Verbmobil-Report, DFKI Saarbrücken, 1996.

3. Alexandersson, J./ Reithinger, N./ Maier, E.: „Insights into the Dialogue Processing of Verbmobil“. Submitted to ANLP-97, Washington, D.C.

4. Geist, A. / Benuelin, A. / Dongarra, J. / Jiang, W. / Manchek, R. / Sunderam, V.: “PVM 3 User’s Guide and Reference Manual”. Oak Ridge National Laboratory , Oak Ridge, Tennesee 1994.

5. Amtrup, J. W./ Benra, J.: „Communication in large distributed AI-Systems for Natural Language Processing“. Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark 1996.

6. Bub, T./ Schwinn, J.: „VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System“. Proceedings of the ICSLP, University of Delaware, 1996.