VERBMOBIL: A SPEECH–TO–SPEECH TRANSLATION SYSTEM*

Damir Ćavar & Wolfgang Menzel

Universität Hamburg, FB Informatik, AB NatS
Vogt-Kölln-Str. 30, D-22527 Hamburg
tel: +49–40–5494 2522, fax: +49–40–5494 2515
e-mail: cavar@informatik.uni-hamburg.de

ABSTRACT

Verbmobil is a speech-to-speech translation project that involves about 29 partners in 3 countries. It started in 1993 and went into its second phase at the beginning of 1997. This phase will last until the end of the year 2000. The main goal of the project is to develop a translation system for spontaneous speech that allows people who speak different languages (i. e. German, English or Japanese) to arrange appointments, make hotel reservations, or get travel information. This paper describes the basic goals of the Verbmobil project, the architecture of the system, and the evaluation efforts made.

1 THE VERBMOBIL SYSTEM

Verbmobil (VM) is an interdisciplinary project in the domain of natural language spontaneous speech-to-speech translation. It is a dialog translation system which aims at facilitating the communication with partners that are native speakers of German, English or Japanese, with limited or no knowledge of the other languages, in specific domains.

1.1 The task

A speech-to-speech translation system is confronted with the task to process spontaneous speech that is not prepared or conceptually well organized. Spontaneous speech contains sentences that are usually judged to be ungrammatical. It consists of a high number of discontinuities, ellipses, corrections and parentheses, as well as non-linguistic elements, like for example breathing noisily, clearing one's throat. Furthermore, spoken language doesn't make use of punctuation.

A system that is based on the written language and prescriptive grammars would thus not be able to process input like "*Well, let me see, mhm, on Monday, hmm, no on Friday, maybe at one, oh no, better at two p. m.*", that represents the majority of spontaneous inputs.

VM makes use of statistical methods and linguistic knowledge to process spontaneous speech in a robust way, being insensitive to errors in the input. Prosodic information, like for example intonation contours and prosodic boundaries, is used for disambiguation. An input like "the-Friday-is-OK-for-you" can only be disambiguated if the prosodic information is available, i. e. this input could either be affirmative ("The Friday is OK for you.") or interrogative ("The Friday is OK for you?"), and would be therefore translated in different ways.

Furthermore, world knowledge and context or dialog specific knowledge plays a major role in the transfer process from one language to the other.

Finally, speech synthesis plays an important role in a speech-to-speech system. For example, the input of a female user should not be translated and synthesized using a male voice. It is therefore important not only to generate an output that has a well formed intonation contour and that sounds "natural", but also to transfer the basic characteristics of the users voice.

1.2 Historical overview

The first integrated system (VM-I) was presented 1995 at the CeBIT in Hanover. This first demonstrator was able to recognize German input in the domain of appointment scheduling, to translate it, and to give the English output.

The demonstrator presented 1997 at the CeBIT also recognizes Japanese input, and translates it into English. Furthermore it is able to process German dialogs of clarification with the user.

During the first phase, the following results were realized in VM-I[2]: a. the recognition of spontaneous speech for German, Japanese and English; b. a lexicon with 2500 words for the translation direction German to English; c. a speaker adaptive system with a

*This research was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 101 A/O. The responsibility for the content of this paper lies with the authors.

speaker independent kernel; d. a basis grammar of German for spontaneous speech, based on linguistic knowledge, with deep and shallow semantic analysis; e. spoken dialogs for clarification between the user and the VM-system in situations with recognition or translation problems; f. semantic transfer for German to English and Japanese to English; g. speech generation for English, and for German paraphrases; h. over 70% approximately correct translations in the end-to-end evaluation in the domain of appointment scheduling; i. pure processing time $< 6 *$ real time (length of the speech signal).

1.3 The second phase

The second phase concentrates on robust and direct translations of spontaneous speech for the language pairs German and English, and German and Japanese. In the second phase, VM will be implemented on a central speech server that can be used over ISDN-telephone lines, ATM-based telecooperation services, or GSM mobile telephones. The server will be able to identify the language used at each input channel, and to recognize, translate and generate the relevant output.
Compared with VM-I, VM-II will have the following additional specifications:

- Multi lingual: It is supposed to translate spontaneous speech in different languages.
- Multi functional: It should be quickly adaptable to different dialog domains.
- Multi medial: It should offer translation aids for international multimedia applications.
- Mobile: With the use of language servers it should also be usable with mobile telephones.
- Multi-party functionality: It should be usable not only in dialog situations, but also in telecooperation systems with many dialog partners.

1.3.1 Financing the project

The first phase of the project (1993–1996) was financed with 64,9 million German marks (DM) by the German Federal Ministry of Education, Science, Research and Technology (BMBF). In addition, 31 million DM were contributed by partners from industry. The second phase of the project (1997–2000) is financed with 50,2 million DM by the BMBF, and with 20,4 million DM by the industrial partners [2].

2 THE ARCHITECTURE OF VERBMOBIL

VM contains basically the following components. Due to time limits, a more detailed description of the different components implemented in the VM system will be given in the final version of this paper.

2.1 Speech recognition

The speech recognizers used in VM are statistical Acoustic input is
VM-I included for the first time a real time speaker independent recognizer for spontaneous speech with a high recognition rate.
VMI was able to process extremely long input sequences (sentences) that are used in negotiation dialogs, as opposed to command input in speech controled systems.

2.2 Prosodic segmentation

VM is the only system that uses prosodic information for translation at different levels of processing, e. g. syntactic, semantic, shallow or stochastic processing. Furthermore, prosodic segmentation introduces information about sentence boundaries that facilitates the syntactic processing for about 92%, and the disambiguation process for about 96%.[4]

2.3 Shallow analysis and translation

The shallow translation splits up a word sequence into communication units. The meaning of the communication units is offered in the translation, independent of the exact linear sequence these units in the original language.

2.4 Stochastic translation

Stochastic translation is based on a bilingual corpus and a statistical model that is trained on this corpus. The model predicts for a word sequence of the input language a word sequence of the target language.

2.5 Deep analysis and translation

In the component of deep translation, the input word sequence is analyzed syntactically and semantically. The linguistic structures are then transfered to the target language.

2.6 Example-based translation

The component for example based translation stores all sentences or parts of sentences in a database. Input sentences are

2.7 Generation

VM includes a very efficient generator that is based on a reversible HPSG-grammar for English. The mean generation time of this generator for one sentence is

0,7 seconds.

3 EVALUATION

The VM system is evaluated every six months. The evaluation of VM-II concentrates on the user perspective. “Real users” are involved in examining the systems properties and usability for arranging appointments or journeys. Available measures from such experiments is the dialog success rate, and the time consumption.[6]
Furthermore, VM-II uses a hierarchical evaluation method, that also involves evaluation of single modules, as well as a turn-end-to-end evaluation. The quality of the input and the output is evaluated on a linguistic basis, analyzing not only the information transmition rate in the transfer process, but also the syntactic and acoustic quality of the input and the output.[1]
Due to time limits, a more detailed description of the evaluation methods can only be presented in the final version of this paper.

4 REFERENCES

[1] Ćavar, D. 1998. *End-to-End Evaluation des Verbmobil Herbstdemonstrators 1998.* Mscr. University of Hamburg.

[2] Karger, R. & W. Wahlster. 1997. Verbmobil: Multilinguale Verarbeitung von Spontansprache. *Künstliche Intelligenz* 4.97, 41–45.

[3] Menzel, W. & J. J. Quanz. 1997. Linguistische Verarbeitung im Maschinellen Dolmetschen: Syntax, Semantik, Transfer. *Künstliche Intelligenz* 4.97, 19-25.

[4] Nöth, E. et al. 1997. Spracherkennung und Prosodie. *Künstliche Intelligenz* 4.97, 14-19.

[5] Wahlster, W. 1993. Verbmobil: Translation of face–to–face dialogs. In *MT Summit IV*, Kobe, Japan.

[6] Jost, U. & W. Menzel. 1998. *Evaluation in Verbmobil.* Mscr. University of Hamburg.