In Proceedings of the Verbmobil TP13 Workshop, Hamburg, Germany. October 18-20, 1995.

NATURAL SPEECH PROCESSING IN PRACTICE: EXPERIENCES WITH THE VERBMOBIL / JANUS-2 SYSTEM

A.Waibel *M.Finke* *D.Gates* *M.Gavalda* *P.Geutner* *T.Kemp* *A.Lavie* *A.McNair*
L.Mayfield *M.Maier* *I.Rogina* *K.Shima* *T.Sloboda* *M.Woszczyna* *T.Zeppenfeld* *P.Zhan*

Interactive Systems Laboratories
University of Karlsruhe, 76131 Karlsruhe, Germany, and
Carnegie Mellon University, USA

ABSTRACT

In speech to speech translation systems for spontaneous input, a variety of problems has to be solved. We introduce the VERBMOBIL / JANUS-2 system and report on our experiences with it, focusing our attention to the problems inherent to the recognition and the parsing of spontaneous speech.

1. INTRODUCTION

The field of natural speech processing has drawn much attention in the last years. However, especially in the domain of speech translation, only very few experimental systems exist that allow the validation of theoretical developments. Between the University of Karlsruhe and our partner laboratory at Carnegie Mellon, Pittsburgh, we have combined the efforts in VERBMOBIL spontaneous german speech recognition with the translation modules of JANUS-2. JANUS-2 [1] is a speech to speech translation system which takes spontaneous speech in German, English or Spanish as input and translates into either English, German, Spanish or Japanese. Our system has provided ample opportunity to study the difficulties that have to be faced when a speech to speech translation system is built. In this paper we try to share some of our experience.

2. SYSTEM OVERVIEW

The system can be divided into three main parts: speech recognition, parsing, and generation in the target language. Each of the main parts consists of a language independent main engine and language dependent configuration data. The system architecture is depicted in figure 1.

In the parser module, three alternative parsers have been employed: the GLR* skipping parser [3], the PHOENIX concept-based parser [4], and the neural net based PARSEC [5]. In this paper, we will focus our attention to the PHOENIX parser.

3. SPEECH RECOGNITION OF SPONTANEOUS SPEECH

There are a number of topics that have to be adressed when dealing with spontaneous human speech. Most of them pose a problem to all three of the basic modules of a speech to speech translation system. However, the response of the system is dominated by the reaction of the speech recognition component. In the following subsections, we therefore list some of the problems inherent to spontaneous speech

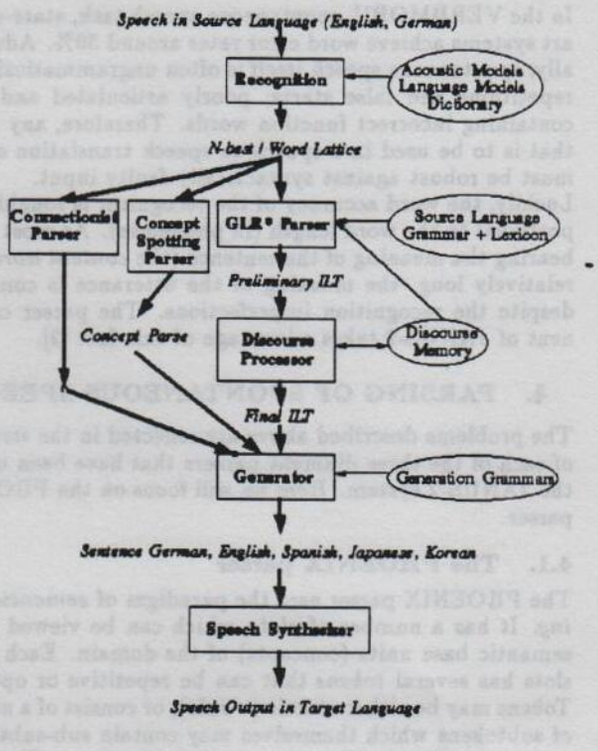

Figure 1. System Architecture

and how they are treated by the speech recognizer of VERBMOBIL / JANUS-2.

3.1. Noise

In spontaneous speech there is a large amount of non-speech noise mixed into the speech data. This noise on the one hand includes *external* noise, like telephone ringing, door slamming and other such events. On the other hand, there is a considerable amount of *human* 'noise', i.e. breathing into the microphone, coughing, or the pause-filling 'mmmmm'.

The VERBMOBIL speech recognition engine introduces specialized *noise models* [6] for both external and human noises. These noise models are added to the recognition dictionary and are inserted into the output of the recognizer like regular words. The use of the information contained in

the noise models for parsing has still to be evaluated.

3.2. The new word problem

Even very large dictionaries cannot cover the full range of a language. This problem is particularly predominant in spontaneous speech. If a word that is not in the recognizer's dictionary is uttered, there will *always* be a recognition error and one or several phonetically similar words will be inserted instead. Recent work in VERBMOBIL [7] aims at locating unknown words and hypothesizing a special token UNKNOWN when such a word has been uttered. However, much work has still to be done to improve the performance of unknown word spotters.

3.3. Recognition errors

In the VERBMOBIL spontaneous speech task, state-of-the-art systems achieve word error rates around 30%. Additionally, spontaneous speech itself is often ungrammatical, with repetitions and false starts; poorly articulated and often containing incorrect function words. Therefore, any parser that is to be used in a speech to speech translation system must be robust against syntactically faulty input.
Luckily, the word accuracy of the recognizer is roughly proportional to the word length (in phonemes). As most words bearing the meaning of the sentence (the *content words*) are relatively long, the meaning of the utterance is conserved despite the recognition imperfections. The parser component of JANUS-2 takes advantage of this fact [2].

4. PARSING OF SPONTANEOUS SPEECH

The problems described above are reflected in the structure of each of the three different parsers that have been used in the JANUS-2 system. Here we will focus on the PHOENIX parser.

4.1. The PHOENIX parser

The PHOENIX parser uses the paradigm of *semantic parsing*. It has a number of *slots*, which can be viewed as the semantic base units (concepts) of the domain. Each of the slots has several *tokens* that can be repetitive or optional. Tokens may be either atomic (words), or consist of a number of subtokens which themselves may contain sub-subtokens, and so on. Slots can be seen as top-level tokens. The topology of a slot thus defines a semantic grammar for a semantic base unit, e.g. a suggestion or an agreement.

The parser matches as much of the input utterance as possible into the tokens of a slot. If a word in the input does not match a token, the slot parse fails. The output of the parser thus is a sequence of successful slot parses. In case of more than one valid parse, the interpretation with the most input words matched is chosen. If there is still ambiguity, the interpretation with the fewest slots is used.

If function words are defined 'optional' in the semantic grammar, the parser will concentrate on the content words of the input. This is a desirable feature, as the recognition error rate for the short function words is higher than for the longer content words.

The end-to-end performance of the system is 44% when translating from English to German.

Note that using the PHOENIX parser, about 70% of the errors were caused by errors of the speech recognizer, 25% were due to coverage problems of the grammar, and only 5% were caused by syntactic ambiguity of the underlying sentence.

A more detailed description of the PHOENIX parser can be found in [8].

5. CONCLUSION

Robust parsing is a key issue in speech to speech translation. As spontaneous speech is often ungrammatical and recognition errors are frequent on spontaneous speech, the use of syntactic parsing techniques is problematic. We have shown that *semantic parsing* can overcome some of the difficulties. A semantic parser could, for example, serve as a powerful backoff component if a more sophisticated syntactic / semantic parsing scheme could not parse the input.

6. ACKNOWLEDGEMENTS

The research described in this paper was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technologie (BMBF) as a part of the VERBMOBIL project. We acknowledge the BMBF support that made this research possible. The views and conclusions contained in this document are those of the authors.

REFERENCES

[1] B. Suhm, P. Geutner, T. Kemp, A. Lavie, L. Mayfield, A. McNair, I. Rogina, T. Sloboda, W. Ward, M Woszczyna, A. Waibel, *JANUS-II: Towards Multilingual Spoken Language Translation*, in Proceedings of the ARPA Spoken Language Technology Workshop, Austin, TX, 1995.

[2] L. Levin, O. Glickman, Y. Qu, D. Gates, A. Lavie, C.P. Rose, C. Van Ess-Dykema, and A. Waibel, *Using Context in Machine Translation of Spoken Language*, in Proceedings of the Theoretical and Methodical Issues in Machine Translation Conference, Leuven, Belgium, July 1995

[3] A. Lavie, M. Tomita, *GLR* - An Efficient Noise-skipping Parsing Algorithm for Context-free Grammars*, Proc. of the Third International Workshop on Parsing Technologies, pp. 123-134

[4] W. Ward, *Understanding Spontaneous Speech: The Phoenix System*, in Proc. ICASSP 1991, vol. 1, pp. 365-367

[5] F.D.Buø, T. Polzin, A. Waibel, *Learning Complex Output Representations in Connectionist Parsing of Spoken Language*, Proc. ICASSP 1994, vol. 1, p. 365 ff.

[6] T. Schultz and I. Rogina, *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition*, Proc. ICASSP 1995, vol 1, pp 293-296

[7] P. Fetter, F. Class, U. Haiber, A. Kaltenmeier, U. Kilian, P. Regel-Brietzmann, *Detection of unknown words in spontaneous speech*, Proc. EUROSPEECH 1995

[8] L. Mayfield, M. Gavaldà, W. Ward, A. Waibel, *Concept-based speech translation*, in Proc. ICASSP 1995, vol. 1, pp 97-100