An Overview of the Slovenian Spoken Dialog System

Ivo Ipšić[1] and Nikola Pavešić[2]

[1]Faculty of Philosophy, University of Rijeka, Croatia
[2]Faculty of Electrical Engineering, University of Ljubljana, Slovenia

In the paper we present the modules of the Slovenian spoken dialog system, developed within the joint project in multilingual speech recognition and understanding "Spoken Queries in European Languages" (SQEL-Copernicus-1634). The system can handle spontaneous speech and provide the user with correct information in the domain of air flight information retrieval. The major modules of the system perform word recognition, linguistic analysis, dialog management and speech synthesis. Some results with respect to word accuracy, semantic accuracy and dialog success rate are given, too.

Keywords: continuous speech recognition, linguistic analysis, dialog management, speech synthesis, spoken dialog system.

1. Introduction

Spoken dialog systems are computer systems that people can call via a telephone network to obtain information without help of a human operator. They handle spontaneous speech and perform a dialog with users using natural spoken language. The Slovenian spoken dialog system has been developed within the joint project in multilingual speech recognition and understanding *Spoken Queries in European Languages* (SQEL-Copernicus-1634), which resulted in a multilingual and multifunctional system, capable of having a dialog with a user in one of the four European languages (German, Slovenian, Czech and Slovak) about a task-oriented topic [1].

Generally, the development of a spoken dialog system concerns solutions to speech recognition problems as well as to speech understanding and human machine interaction problems. The major problems in the development of a continuous speech understanding systems arise due to the nature of the spoken language: there are no clear boundaries between words, since the phonetic beginning and ending of words are influenced by neighboring words; additionally, variability in speech between different speakers can be noticed, and the speech signal may be affected by noise. To avoid these difficulties, spoken dialog systems are usually limited by different constraints: the vocabulary size is about one thousand words, the communication domain is task-oriented, and the sentence structure is usually limited by a simple grammar.

Figure 1. shows the architecture of a spoken dialog information retrieval system. The major modules perform word recognition, linguistic analysis, dialog management and speech synthesis. User utterances are first digitized and transformed into a sequence of speech signal feature vectors. The sequence is passed to the word recognition module, which generates hypotheses of spoken word chains. The recognized words are handed to the linguistic module, which extracts a set of semantic concepts from the recognized words. These concepts are passed to the dialog manager. The dialog manager performs a database query if enough parameters are available. According to the dialog history and dialog strategy the user is asked to confirm the parameters or additional parameters are requested. Database query results are transformed into sentences and, using the speech synthesis module, played to the user over the telephone line.

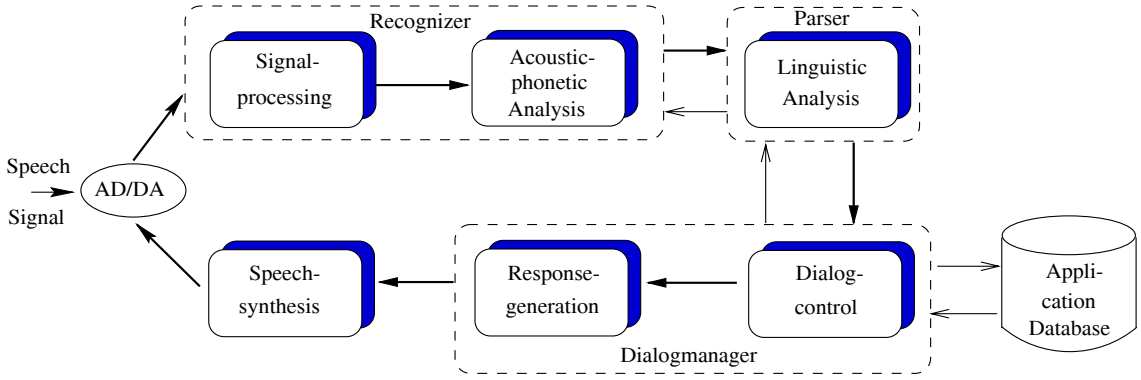

Fig. 1. Architecture of a spoken dialog information retrieval system.

The information system being developed for Slovenian speech is used for air flight information retrieval. Architecturally, it is based on the Erlangen Train Time Table Inquiry System EVAR [2].

In the subsequent sections we describe the modules of the Slovenian spoken dialog system. In Section 2 we present the Slovenian speech database. In Section 3 we describe the word recognition module and address the Slovenian acoustic and language modelling problems. In Section 4 methods and procedures for Slovenian linguistic analysis are presented. The dialog manager and different dialog strategies the system can perform are described in Section 5. Slovenian speech synthesis is briefly described in Section 6. In Section 7 conclusions are summarized.

2. The Slovenian Speech Database

The first step in spoken dialog system design is collection of speech material, which is used for dialog modelling as well as for building statistical models for the word recognizer. We have collected recordings of dialogs between anonymous clients and telephone operators at the Adria Airways information service. From these dialogs we created the Slovenian speech database GOPOLIS of 5000 sentences, which was read by 50 speakers [6]. The sentences were grouped into four groups: introductory and concluding parts of the dialogs, central parts of the dialogs concerning the information domain, questions that would be redirected to another address and phrases consisting of words determining time and date. The read speech database was used for the creation of acoustic and language models as well as for testing of the parser and the dialog manager. We tested the separate modules with microphone and telephone quality speech. The overall system was tested with cooperative users and dialog determination and success rate was estimated.

3. Word Recognition

The task of the word recognition module in a spoken dialog system is to generate word hypotheses that best fit the input utterance. Given a sequence of acoustic units, which can be a sequence $\mathbf{X} = \mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_T$ of T speech signal feature vectors, we have to determine the sequence of spoken words $\mathbf{w} = w_1, w_2, \ldots, w_n$. The most probable word sequence w^* is given by:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\, P(\mathbf{X} \mid \mathbf{w}) \cdot P(\mathbf{w}).$$

The first term in the equation describes the conditional probability of the acoustic units, given the word sequence. We refer to the estimation of $P(\mathbf{X} \mid \mathbf{w})$ as the acoustic modelling problem. The second term of the equation defines the a priori probability of the word sequence $\mathbf{w}$, given by the language model. This probability is independent of the acoustic units, and can be estimated from an application-dependent text corpus. The language model constrains the possible word sequence and incorporates knowledge about syntax and semantics of the spoken language into the recognition procedure.

3.1. Acoustic Modelling

The speech signal is sampled at 16 kHz with 16 bit and mel-cepstrum features and their derivatives are computed every 10 ms. The speech signal vectors are transformed into symbols using a soft vector quantization technique. The recognizer is based on semi-continuous hidden Markov models of context dependent phone units-polyphone units. A polyphone consists of a phone with arbitrary length of left and right context of phones, and it is determined from the training database. The only criterion to form a polyphone is the minimal number of its occurrences in the words of the training sentences. To define the context dependent units we have proposed a set of 33 Slovenian phones and their HMM models, which give the highest recognition accuracy [9]. To train the polyphone models we use the ISADORA system [7]. The 824 words are modelled with 2086 polyphone models of different length.

3.2. Language Modelling

The language model determines the a priori probability of a word sequence $\mathbf{w}$, where $P(\mathbf{w})$ is approximated with conditional probabilities, i.e. a word is conditioned by all its predecessors. If only $n-1$ previous words are used such language models are called n-gram models [7]. Since many of the possible word sequences do not occur in the training corpora it is necessary to resolve the problem of missing word sequences. One approach to the problem of missing word pairs is the use of word categories. Words from the recognition vocabulary are assigned to word categories, and so the number of parameters of the language model can be drastically reduced. We defined 127 word categories to classify the words from the Slovenian recognition vocabulary. The categories group words with the same grammatical and semantical characteristics. The Slovenian language models were trained on 10000 sentences comprising 824 different vocabulary words from the flight information domain.

3.3. Recognition

For word recognition word models are constructed by concatenation of phoneme models, which again are obtained by concatenating the polyphone models according to the pronunciation lexicon [8, 9]. All word models are concatenated in parallel and form a single Hidden Markov Model, which is represented by a huge network of nodes. The analysis of an unknown observation sequence is performed by the Viterbi algorithm, producing the maximum a posteriori state sequence of the model with respect to the observed input vectors. Knowing the state sequence of the HMM we can decode the input sequence and transform it into a string of words. Because of the large number of states which have to be considered when computing the Viterbi alignment, a state pruning technique has to be used to reduce the size of the search space. We use the Viterbi beam-search technique which expands the search only to states whose probability falls within a specified beam. The probability of reaching a state in the search procedure cannot fall short of the maximum probability by more than a predefined ratio. During the forward search in the HMM N best word sequences are generated using acoustic models and a bigram language model. The hypothesized word sequences are then verified using a trigram language model. The result is the most probable word sequence. The recognizer has been evaluated on microphone signals as well as on telephone speech. The speaker-independent word recognition accuracy is over 90%, while the word accuracy for telephone quality speech signals is 76% on average.

4. Linguistic Analysis

Input to the linguistic module is the recognized word sequence; output of the linguistic module is its semantic interpretation in the Semantic Interface Language (SIL) [5]. SIL was defined in the Sundial project [3] for interfacing different parsers with the dialog manager and for knowledge representation within the dialog manager. The Slovenian linguistic analysis module produces a SIL structure from the recognized utterance. The linguistic analysis module consists of a Slovenian parser and a domain-dependent linguistic knowledge base. Besides the known problems which occur during the design of a linguistic analysis system additional problems arose due to the nature of the

Slovenian language: a large number of inflected word forms and a rather free word order. To overcome these problems, we developed a robust semantic-driven parser [10], which extracts the most important information from the recognized sentences. First certain keywords are combined into word phrases (for example temporal expressions are grouped into one phrase), and then compared to predefined templates. The templates consist of semantic classes relevant to the timetable inquiry task, concerning data on flights, departure and arrival places and time, airlines, etc. The recognized sentences are parsed in three different steps which perform:

- parsing of temporal expressions,
- parsing of simple noun words. Their meaning is often deduced from a simple semantic category which they belong to, and from their position in the sentence. This method is used for departure/arrival city determination.
- The rest of the sentence is parsed using a very simple parser that tries to locate as many keywords as possible. These keywords are used in further processing to determine the sentence type.

The semantic-driven parser is based on the Definite Clause Grammar formalism, and is implemented in Prolog [11]. It can handle also ungrammatical and colloquial expressions, but there are phrases, which the linguistic analysis module cannot parse correctly. These are:

- sentences where the time phrase is divided throughout the sentence,
- sentences where determination of arrival or departure town can fail since there is no difference between the first and fourth case of a noun and
- some complex time expressions.

The output of the Slovenian parser is the SIL structure of the recognized user sentence, and it contains only application relevant information. The example shows the SIL structure of the user sentence *Tomorrow in the morning I would like to fly to London.*:

```
O = [[id:1456,semantics:
          [id:1457,type:dbflight,
           date:210100,
           goalcity:london,
           sourcetime:[300,1000]]]]
```

The user intention spoken in the sentence is described with the *dbflight* SIL structure, which contains information about travel date and time, and the place of arrival. The time phrase *in the morning* is mapped onto the interval 3h00-10h00.

5. Dialog Management

Figure 2. shows the structure of the dialog manager. The dialog manager takes the semantic representation of the user utterance and performs the interpretation within the current dialog context and generates system answers. According to the application domain, the dialog manager has to answer user inquires or request

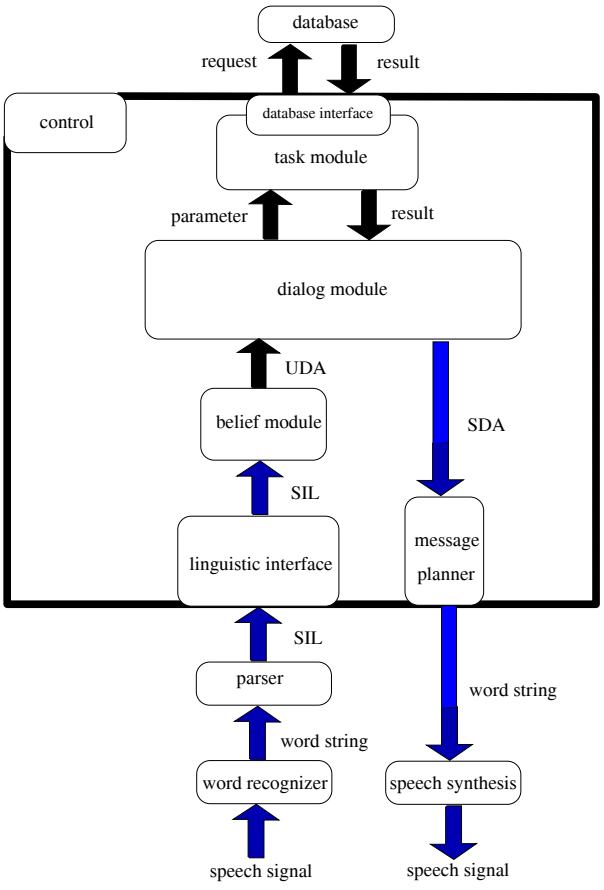

Fig. 2. Structure of the Dialog module in the Slovenian spoken dialog information retrieval system.

additional information needed for a successful database query. The dialog manager used for the Slovenian dialog system, stems from the German EVAR system, and is language- and application-independent. The dialog manager consists of a number of communicating subcomponents [4](Figure 2.):

- The linguistic interface, which enables the input from the parser, extracts the information which is necessary for a database query.
- The belief module is responsible for interpreting user utterances in the current dialog context as well as for resolving ambiguous or inconsistent semantic representations.
- The dialog module is the central part of the Dialog manager. It is responsible for the pragmatic interpretation and for planning of system utterances. It has to keep track of the user dialog acts (UDA), generated by the belief module, and it has to produce system dialog acts (SDA).
- The task module is the interface between the dialog manager and the task-dependent database and is responsible for handling a database query.
- The message planer is responsible for message generation.

The adaptations for a new language and a new application have been done within the task module and the message planer. For the Slovenian dialog system the flight timetable, provided by the Slovenian airline Adria Airways is used. The timetable, which is available on the WWW, is transformed into Prolog facts representing the timetable data.

In the Message Generation Module the SIL representations of the system answers were converted to (Slovenian) text via sentence tabloids. These are system sentences which either ask users for additional parameters or confirmations and sentences with database query results.

To get a feel for the capabilities of a spoken dialog system we show below an example of a typical successful dialog where a user is asking the system about flight connections.

system: *Hello, this is the Adria Airways Air Flight Information System. How can I help you?*
user: *tomorrow afternoon I would like to fly to London*
system: *Do you want to fly on Wednesday the 23rd of April between 12 and 19? Do you want to fly to London? Where do you want to leave from?*
user: *yes from Ljubljana*
system: *Do you want to leave from Ljubljana?*
user: *yes*
system: *To London you fly on Wednesday the 23rd April from Ljubljana at 14:10, arrival at London at 15:25, I repeat, departure at 14:10, arrival at 15:25. Shall I repeat?*
user: *no*
system: *Do You need any additional information?*
user: *no thanks*
system: *Goodbye*

5.1. Dialog Strategies

To improve the clarification dialogs different dialog strategies are defined in the dialog module. They define the initiative in the dialog acts and different confirmation strategies of the recognized parameters, which are needed for a database query. These strategies are: separate confirmations of SIL parameters, confirmation alone, confirmation of SIL parameters and dialog initiative and no confirmation [5]. On the example of the recognized user sentence *"Tomorrow in the morning I would like to fly from Ljubljana to London."* different dialog strategies are demonstrated:

- Separate confirmations:

  system: You want to fly to London?
  user: Yes.
  system: You want to leave from Ljubljana?
  user: Yes.
  system: When do you want to travel?

- Alone confirmations:

  system: You want to fly from Ljubljana to London?
  user: Yes.
  system: When do you want to travel?

- Confirmation of parameters and dialog initiative:

  system: When do you want to travel from Ljubljana to London?

- No confirmation of parameters:

  system: When do you want to travel?

Depending on the dialog strategies the dialogs can be long and boring and they can be short but unfortunately unsuccessful. The experiments with the Slovenian spoken dialog system show that the "no confirmation" strategy leads to more unsuccessful dialog ends. The reason lies in the word recognition rate, which is about 80%. Best dialog success rates are achieved with "separate confirmation" and "confirmation alone" dialog strategies. Confirmation of recognized parameters or rejection of wrong recognized parts of sentences leads to more successful dialogs. The ideal solution would be a dynamic dialog strategy, which would change depending on the recognition and linguistic analysis success rates.

6. Speech-synthesis

The generated system answers are transformed to speech signals. This can be done with a simple concatenation technique of prerecorded time signals for isolated words. Another possibility is the use of synthesized speech. A system for Slovenian synthesis was developed [12]. The Slovenian synthesizer uses the concatenating diphone synthesis technique TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) [13]. The TD-PSOLA scheme enables pitch and duration transformations directly on the waveform, at least for moderate ranges of prosodic modifications without considerably affecting the quality of synthesized speech.

7. Conclusion

In the paper we overviewed the Slovenian spoken dialog system, architecturally based on the German spoken dialog system EVAR. The main necessary adaptations for Slovenian language and for a different information retrieval task were: definition of subword units, modifications of language model, parser and dialog manager as well as a new implementation of the text-to-speech module (in our case).

Experiments with typed and spoken input have shown that database queries can provide the users with the desired information. Besides word recognition rates and linguistic analysis success rates, a major role in the dialog success rate is played by the dialog strategy. We have presented different dialog strategies in the dialog manager, and shown which strategy leads to best dialog success rate. These results are appropriate for the Slovenian spoken dialog system, which enables a dialog with a user over the telephone line. So far we have collected a number of spontaneous spoken dialogs over the telephone. The results are encouraging and show that in about 50% of the dialogs the users get the desired information.

References

[1] M. ARETOULAKI, S. HARBECK, F. GALLWITZ, E. NOETH, H. NIEMANN, J. IVANECKY, I. IPŠIĆ, N. PAVEŠIĆ, V. MATOUSEK, *SQEL: A Multilingual and Multifunctional Dialog System*. In *ICSLP'98 The 5th International Conference on Spoken Language Processsing*, Volume 3, pp. 855-858, 1998.

[2] F. GALLWITZ, M. ARETOULAKI, M. BOROS, J. HAAS, S. HARBECK, R. GRUBER, H. NIEMANN, E. NOETH, *The Erlangen Spoken Dialogue System EVAR: A State-of-the-Art Information Retrieval System*. In Proc. International Symposium on Spoken Dialogue (ISSD 98), Sydney, pp. 19-26, 1998.

[3] J. PECKHAM, *Speech Understanding and Dialogue over the Telephone: an Overview of Progress in the SUNDIAL Project*. In *Proc. European Conf. on Speech Technology*, volume 3, pages 1469-1472, 1991.

[4] W. ECKERT, *Customizing the Erlangen Dialog Manager*. Technical report, Lehrstuhl für Mustererkennung, FAU Erlangen-Nürnberg, 1996.

[5] W. ECKERT, *Gesprochener Mensch-Maschine-Dialog*, Ph. D. Thesis, Universität Erlangen-Nürnberg, 1996.

[6] S. DOBRIŠEK, J. GROS, F. MIHELIČ, N. PAVEŠIĆ, *Recording and Labelling of the GOPOLIS Slovenian Speech Database*, First International Conference on Language Resources and Evaluation, eds.: A. Rubio, N. Gallardo, R. Castro, A. Tejada, May 1998, Granada, Spain, Vol. II, pp. 1089-1096.

[7] E.G. SCHUKAT-TALAMAZZINI, *Automatische Spracherkennung*, Vieweg, Braunschweig, 1995.

[8] N. PAVEŠIĆ, *Continuous Speech Recognition by a Network of Hidden Markov Models*, Journal of Computing and Information Technology, Vol. 3, No. 3, pp. 193-205, 1996.

[9] I. IPŠIĆ, F. MIHELIČ, N. PAVEŠIĆ, E. NOETH, *Slovenian Word Recognition*, Proceedings of IAPR Workshop on "Speech and Image Understanding", N. Pavešić, H. Niemann, S. Kovačič and F. Mihelič eds., Ljubljana, Slovenia, pp. 87-96, 1996.

[10] K. PEPELNJAK, F. MIHELIČ, N. PAVEŠIĆ, *Semantic Decomposition of Sentences in the System Supporting Flight Services,* Journal of Computing and Information Technology, Vol. 4, No. 1, 1996, pp. 17-24.

[11] F. PEREIRA, D. H. D. WARREN, *Definite Clause Grammar for language analysis-a survey of the formalism and a comparison with augmented transition networks*, Artificial Intelligence 13, pp. 231-278, 1980.

[12] J. GROS, N. PAVEŠIĆ, F. MIHELIČ, *Text-to-Speech Synthesis: A complete system for the Slovenian Language*, Journal of Computing and Information Technology, Vol. 5, No. 1, 1997, pp. 11-19.

[13] E. MOULINES, F. CHARPENTIER, *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*, Speech Communications **9**, pp. 453-467, 1990.

Received: February, 2001
Revised: April, 2002
Accepted: September, 2002

Contact address:
Ivo Ipšić
University of Rijeka
Faculty of Philosophy
Omladinska 14, HR-51000 Rijeka
Croatia
Phone: +385-51-345-046
Fax: +385-51-345-207
e-mail: ivo.ipsic@pefri.hr

Nikola Pavešić,
University of Ljubljana,
Faculty of Electrical Engineering,
Tržaška 25, SI-1000 Ljubljana
Slovenia
Phone: +386-1-4768-315
Fax: +386-1-4768-316
e-mail: nikola.pavesic@fe.uni-lj.si

IVO IPŠIĆ was born in 1963. He received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana in 1988, 1991 and 1996, respectively. From 1988-1998 he was a staff member of the Laboratory for Artificial Perception, at the Faculty of Electrical Engineering, University of Ljubljana. From 1996-1998 he was working as a research associate on the project Spoken Queries in European Languages (SQEL-Copernicus-1634). He was a guest researcher for a period of 10 months during the academic year 93/94 and for two months in 1996 at the Friedrich Alexander University of Erlangen-Nürnberg, Department of Computer Science. His work concerned acoustic modelling for continuous speech recognition. Since 1998 he is an assistant professor of computer science at the University of Rijeka. His current research interests belong to the field of multilingual speech recognition and digital signal processing. He is a member of the International Speech Communication Association and the Slovenian Pattern Recognition Society.

NIKOLA PAVEŠIĆ was born in Rijeka, on December 7 1946. After his final exam at the natural science and mathematics department of the II. Grammar School in Rijeka, he studied at the Faculty of Electrical Engineering, University of Ljubljana. There he received the B.S. degree in electronic, the M.S. degree in automatics and Ph.D. degree in electrical engineering in 1970, 1973 and 1976, respectively. He was the recipient of the Mario Osana Award in 1974, the Vratislav Bedjanič Award in 1976, the Boris Kidrič Fund Award in 1982, and the Milan Vidmar Award in 1996. Since 1970 he has been a staff member of the Faculty of Electrical Engineering in Ljubljana, where he is currently a Full Professor of systems, automatics and cybernetics, Head of the Laboratory for Artifical Perception, Systems and Cybernetics, and Vice-chairman of the Department. His research interests include pattern recognition and image processing, speech recognition and understanding, and theory of information. He has authored or co-authored more than 200 papers and 3 books addressing several aspects of the above areas. Dr. Nikola Pavešić is a member of IEEE, The New York Academy of Science, the Slovene Association of Electrical Engineers and Technicians (Meritorious Member), The Slovene Pattern Recognition Society (Founder and first president), and the Slovene Society for Medical and Biological Engineering.