SYSTEM OF AN AUTOMATIC SPEECH RECOGNITION AND SPEECH UNDERSTANDING LINGVO/LASER
(SYSTEM DER AUTOMATISCHEN SPRACHERKENNUNG UND DES SPRACHVERSTEHENS LINGVO/LASER)

Václav Matoušek, Kamil Ekštein, Tomáš Pavelka

Department of Computer Science and Engineering
Faculty of Applied Sciences
University of West Bohemia in Plzeň (Pilsen)
matousek@kiv.zcu.cz

Abstract: The presented paper introduces a scalable modular prototyping concept and methodology framework for rapid development of domain-oriented information retrieval dialogue systems being developed in co-operation of the above mentioned department and the Department of Informatics of the Faculty of Art IV of the University of Regensburg. The system consists of two stand-alone aggregates of software modules and knowledge bases – speech recognition and synthesis software (LASER) and a generic task-independent scalable dialogue system (LINGVO). The article presents the architecture, design, and philosophy of the framework in details and discusses its benefits experienced in development of an experimental dialogue information system (DIS).

1 Introduction

Many research facilities (universities, research institutes, specialized private enterprises) offer various kinds of speech-enabling software in this time. Offered systems usually make it possible to build voice-driven (mostly telephone operated) information retrieval dialogue systems that provide their prospective users with certain information from selected domains, e.g. arrivals and departures of trains, buses, planes, public transportation vehicles at all, opening hours of some institutions and/or shops, currency exchange rates, phone numbers, etc. The developed system has to be able to satisfy successfully requests of its users. As the overall recognition accuracy of even the state-of-the-art systems is still far from being sufficient to work on wide domains, the designers must back up on highly restrictive techniques giving the user in fact no chance to mislead the system – the dialogue then might look like the following:

System:	Welcome to the airplane ticket reservation system. Where do you want to fly from and to?
User:	***from Prague to Berlin***
System:	Your departure city is Prague. Your arrival city is Berlin. A ticket from Prague to Berlin is reserved for you. Is it so okay ?
User:	***yes thank you***
System:	Do you have any other request ?
User:	***no thanks***
System:	Thank you for your interest, have a nice day. Good bye.

This dialogue represents the most appropriate (user friendly) alternative of the dialogue, no additional information, no data corrections are needed. Many users are not able to formulate their requests clearly and definitely, and then more clarification questions must be generated by the system and answered by the user. Such dialogue can look like the following:

System: Welcome to the airplane ticket reservation system. Where do you want to fly from and to?
User: ***Prague***
System: Your departure city is Prague. Is it so okay ?
User: ***yes***
System: To which city do you want to fly?
User: ***Bowlin***
System: Sorry, I did not understand you. Tell me to which city you want to fly, please
User: ***Bowlin***
System: I am sorry, I still do not understand you. You probably used the destination which I do not know or you have mispronounced the city name. Would you be so kind and clearly repeat the city name you want to fly ?
User: ***Berlin***
System: Your arrival city is Berlin. Is it so okay ?
User: ***yes***
System: A ticket from Prague to Berlin is reserved for you. Have you any other request?
User: ***no thank you***
System: Thank you for your interest, have a nice day. Good bye.

Both dialogues above are acceptable for, say, novice user. But, an experienced one would for sure like to “speed up” the dialogue causing the system to get confused and terminate the dialogue immediately by means of some kind of escape utterance. Therefore modern commercial dialogue systems use a full-featured, mature human-computer interaction based on an advanced “barge-in” dialogue strategy [9, 10]. This fact led the team of the Pilsen Department to start the development of a dialogue information system prototyping concept which will make it possible to build voice-driven applications without restrictions in spoken interaction. The fundamental inspiring idea of the design is to extract as much information as possible at any level of the communication, and use it back at the lowest level possible.

2 Language Modeling Considerations

As the phoneme recognition accuracy can hardly exceed some 80 %, the relatively high utterance recognition accuracy (reported about 95 – 97 % in the state-of-the-art systems) grounds in powerful restrictive language modeling which is capable of rejection of incorrect hypotheses (referred to as out-of-grammar hypotheses). In the Czech language the restrictive power of grammar (as well as statistical language models) is significantly debilitated by syntactical properties of the language. At first, the Czech language has free word order [7], i.e. a lot of possible word groupings are acceptable and cannot be considered out-of-grammar. An another property of the Czech language is a full-featured flection – nouns, pronouns, adjectives, and numerals are declined into seven cases for each grammatical number (resulting usually up to 12 different word forms), and verbs are conjugated in a very complex way resulting in a nightmare of 223 different forms of a verb). Both declination and conjugation are (mostly) suffix-based, misrecognized suffix may end up in a completely different meaning of the utterance. Taking the grammatical structure of recognition hypothesis into account may result in rejection of a generally correct hypothesis due to any misrecognized suffix. And thirdly – the language model perplexity rises significantly.

A series of tests with HTK 3.2 toolkit trained with three corpora was carried out at the Laboratory of Intelligent Communication Systems (LICS). The values of best phoneme recognition accuracies achieved there lie between 72.5 % and 82.9 %, word recognition accuracy in the best recognized sentence hypothesis reached 77.6 % by using “standard” grammar, and 96.3 % by using the most restrictive grammar. The most restrictive grammar forces user to announce his or her intention in a very tight manner. Such a language model is disputably applicable for any Czech as it gives no freedom of word ordering which is very natural for us. On the other hand the “standard” grammar covers nearly all possibilities of free word order sentences applicable to express the “fine” sentence meaning. It unfortunately results in a dramatic drop of performance. An another chance to increase the word recognition accuracy is to use permugram language models [14]; their use will be tested in the future.

3 System Design Considerations

As the grammar or statistical language model cannot play its restrictive role in the Czech “speaking” DIS, we decided to derive the restrictions from dialogue course, generally at any level of the dialogue system. To clarify this idea, let us consider the following situation: The system is asking the user “Do you have a credit card ?”. There is very high probability that the user answers either “***yes***” or “***no***”. We examined hundreds of recordings and found only few rare cases when the user faced to a pure yes/no-question replied anything else – if he or she did so, the dialogue was not co-operative at all anyway [7, 15].

The points of a dialogue system design where an appropriate restrictive information can be derived from, can be e.g. the following:

1. **Acoustic Front-End** (signal processing):
   (a) Measuring fundamental voice frequency F_0 can tell whether the speaker is male or female. Such knowledge can be used in (i) acoustic-phonetic decoder to switch to an appropriate set of models (HMM) or neural nets (ANN) trained by men or women respectively; (ii) language modeling to conceal the grammar components for female forms (endings and another gender-specific phenomena).
   (b) Measuring prosodic parameters (e.g. overall loudness) to detect anger or stress can help too switch to a human operator (if available) in due time.
2. **Domain Analysis:** May influence the language modeling knowledge base by means of iteratively narrowing the vocabulary and grammar to the discussed domain plus some escape utterances.
3. **Data Analysis:** Modifies situation modeling knowledge bases to exclude dialogue sequences leading to a query about a fact which is not known to the system or which the system cannot answer for any reason.
4. **Dialogue Manager:** Being the main decisive mechanism of the dialogue system, the dialogue manager is a source of wide set of information – e.g. following dialogue situation can result in considerable restriction of a language model in those branches where the user has lesser freedom of choice and thus possible interaction is predictable according to the dialogue scenario**.**

4 System Architecture Description

Figure 1 shows schematically the architecture of the LASER/LINGVO framework. The whole prototyping concept has been designed to enable applying of modeling restrictions according to knowledge acquired all around the system. Modules and functional units of the system design are described in detail below:

4.1 LASER – Speech Recognizer

1. **Run-time Recorder** controls computer audio device(s) and records an incoming speech into a stream of digital data. It incorporates voice activity detection (VAD) and automatic gain control (ACG) bocks too. Parameter setup (sample rate, quantization parameters) is user adjustable via configuration file and/or command line options.

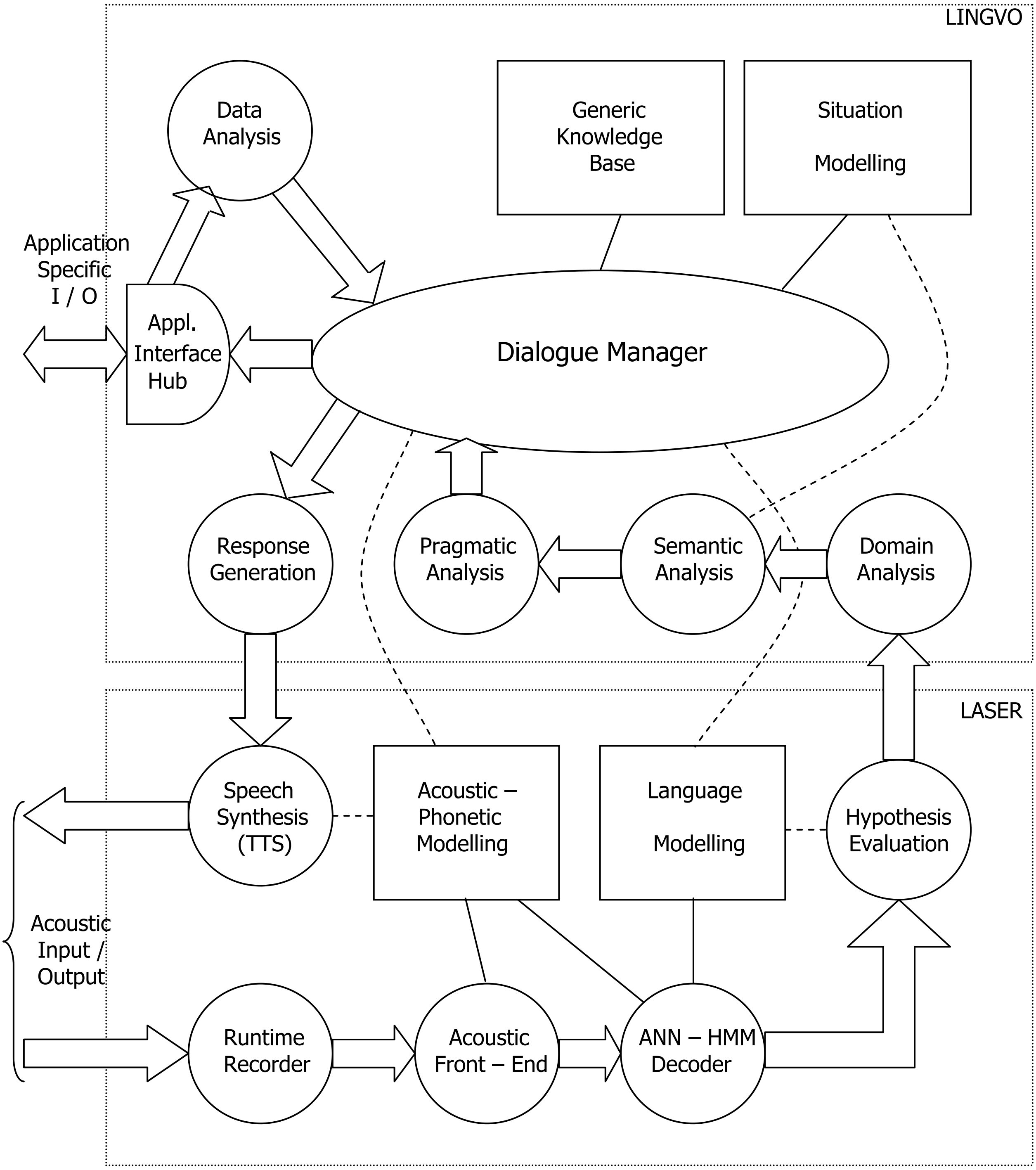

Fig.1: Architecture of the LASER/LINGVO system: Circular elements represent executive modules (routines, programs, software tools], rectangular elements stand for knowledge bases (files, databases, knowledge systems), double arrows show data flow through the system, solid lines connect parts that exploits one another, end dashed lines connect those that share and/or enriches knowledge bases.

2. **Acoustic Front-End** transforms the recorded digitized speech signal into a stream of parametric vectors used for further processing. The content of the parametric vector is user adjustable by a script in SACL/PDL language which defines the exact way to treat the signal. Possible processing options include preemphasis, smoothing, windowing, power spectral density estimates (smoothed), spectral warping (Mels), MFCC, PLP, liftering, mean and deviation normalization, and many others [4].

3. **ANN-HMM Acoustic-Phonetic Decoder** decodes the spoken utterance represented by parametric vectors into phonetic information (series of phonemes represented by transcription alphabet symbols) by means of proposing recognition hypotheses based on acoustic and language modeling. Artificial neural network (namely MLP) estimates aposteriori probabilities of phonetic class assignment for each parametric vector. These values are used as output probabilities b_j in states of HMMs of phoneme-like units [13].

4. **Hypothesis Evaluation** searches the proposed recognition hypotheses in the shape of word lattice and, according to the language modeling knowledge base (and thus information provided by upper level of the design, e.g. dialogue manager), accepts the most probable way(s) through the lattice, i.e. the valid hypothesis.

5. **Speech synthesis (TTS)** shares the acoustic/phonetic knowledge base to produce audible speech output. There are several TTS systems at disposal [6].

6. **Acoustic-Phonetic Modeling Knowledge Base** contains models of acoustic phenomena and their phonetic class assignment in form of numeral parameter sets for HMM (matrices of a_{ij} and b_j, transition and emission probabilities). The knowledge base can be enriched by external knowledge according to the propose design concept of gathering and spreading knowledge: (a) the acoustic front-end module can determine whether the prospective speaker is male or female and cause switching too the appropriate set of acoustic-phonetic models instead of using both two – resulting performance improvement is estimated about 10 %; (b) the same piece of information can come from dialogue manager (as Czech language strongly differentiates grammatical gender).

7. **Language Modeling Knowledge Base** contains language models, i.e. grammar in e.g. Advanced Backus/Naur Form (ABNF), numeral parameter sets of N-gram statistical models, etc. This base can be also strongly influenced by spreading knowledge from upper parts of the design – the information from situation modeling base (via dialogue manager) can suppress grammar branches that will not be used for sure next user's reply. Our contemporary technical solution of this task is a re/generation of the used grammar before each utterance analysis.

4.2 LINGVO – Dialogue System

1. **Domain Analysis:** At this point, a decision about what domain does the utterance belong to is taken. According to such a knowledge, an appropriate situation models are passed to the dialogue manager. Also an off-topic sentence can be identified here and the dialogue manager is consequently alerted to switch to an "escape" scenario. The module is based mainly on the vocabulary and syntax analysis [1, 12].

2. **Semantic analysis** analyses the utterance with the goal to find the meaning of it, i.e. expressed intention of the speaker in the communication towards the system. Semantic analysis is based on microsituation theory and several other semantic formalisms [12]. The method tries to fill in predefined semantic frames using the information contained in the sentence – those frames that are filled more than certain given level are declared valid semantic hypotheses and passed to the next module.

3. **Pragmatic Analysis** verifies whether the semantic hypothesis is accomplishable given the contents of domain-specific databases. Pragmatic analysis also contributes to quantitative formulation of the cooperativeness level between the user and the system. Such information helps to select suitable dialogue strategies within the situation modeling base.

4. **Data Analysis** scans the data produced by controlled applications and returned to the dialogue system through the interface hub. The module is responsible for filtering

singularities from the data and translating the data into semantic frames so that dialogue manager can operate on them.

5. **Interface Hub** ensures communication with controlled applications (subordinated) such a relational databases, system terminals, game engines, etc.
6. **Response Generation** translates filled-in data frames back to human speech in the form of a sequence of phonetic symbols, which is further passed to the speech synthesizer.
7. **Generic Knowledge Base** contains common facts needed to decode incomplete semantic frames or those carrying implicit entries like e.g. local date and time, position of the running system, etc. In the other words it holds a system-specific description of the world.
8. **Situation Modeling Knowledge Base** contains dialogue and subdialogue scenarios derived out of long-lasting research of real human-human dialogues, dialogue templates, and behavioral patterns [15]. This module is a prominent source of knowledge used to restrict the recognizer grammar.

5 Current State of Implementation

The following program units and modules are fully functional:

1. **LASER Recognizer Unit** – it provides the system with either the best recognized sentence hypothesis or N-best hypotheses. The experimental hybrid ANN-HMM decoder [4] may be optionally replaced by HTK/ATK-based decoder. Implemented also as DLL library module, the recognized may be utilized by various speech-enabled applications too.
2. **Domain Analysis**
3. **Interface Hub**
4. **Response Generation**

Interface routines (written in Perl) enable to incorporate executive modules or data from other systems,, e.g. HTK, CSLU Toolkit, or SPEX KIT. The Semantic and Pragmatic Analysis modules and the Dialogue Manager are partially implemented, i.e. they are available in a simple form for testing and display purposes. Still they are not ready as generic full-featured data-driven modules. Currently a co-operative effort is exerted to bind LASER/LINGVO system to SPEX KIT dialogue platform [16].

A complete methodology is prepared for the dialogue modeling – microsituations, dialogue flow, escape strategies, etc. Also several real recorded dialogues were modeled using the methodology to verify its efficiency [8].

Several simple dialogue systems have been developed using the LASER/LINGVO framework: (i) LChess – a chess game controlled by voice interaction; DOD@live – a dialogue system for "Day of Open Doors" at the Department of Computer Science answering questions of our prospective students about the study programs at the department; (iii) CIC (City Information Center) – a municipality dialogue information system providing information about city transportation, opening hours of institutions, etc.

6 Results and Future Work

The way how the prototypes nowadays function (on an isolated Windows-based workstation with an earphone/microphone headset) does not allow an extensive testing under real operating conditions. We performed only a simple test during the above mentioned "Day of Open Doors" when 113 uninitiated students (they were not previously instructed how to speak to the system and what to say) talked to the DOD@live dialogue system prototype. The results were as follows:

Wrong system response	**6.81 %**
Correct system response	**93.19 %**
> Correct hypothesis (A)	59.09 %
> Wrong hypothesis (B)	34.09 %

State (A) means that the recognizer provided the system with correct hypothesis and the system subsequently took an appropriate action (response) so that the user was satisfied. State (B) is a situation when the recognizer provided the system with (partially) incorrect hypothesis but the system was still able to derive the meaning of the utterance and take an appropriate action (response) to satisfy the user.

The weakest point of current LASER/LINGVO implementation state is definitely the semantic and pragmatic analysis as these modules can act as efficient restriction of recognizing hypotheses. Also a rejection mechanism for totally out-of-dialogue hypotheses with high recognition score (Hypothesis Evaluation module) works at disputatious level of accuracy leading the system to "dead ends".

Our future work will be focused towards implementing data-driven algorithms for semantic and pragmatic analysis. Another important branch is to improve the dialogue manager core to (i) handle exceptional dialogue states, (ii) support escape strategies, and (iii) cover wider field of dialogue situations (i.e. make the frame processing more generic). Moreover we would like to incorporate some recently presented NLP techniques suitable for Czech language but unfortunately these are usually too theoretic and too demanding to be implemented in real-time responding system.

7 Conclusion

The information retrieval dialogue system paradigm described in the previous paragraphs has been proposed and built mainly because of the need of a design concept enabling to increase the cooperative performance between a human and a machine. Such result is strongly dependent on the speech recognition and semantic analysis accuracy as these are key components in the process of artificial understanding of speech. The design was penetrated with a fundamental idea of highest possible modeling restrictions so that the decoding algorithms have lesser freedom and thus gaining better results. The need for such a scheme came out of syntactical properties of the Czech language for which both grammar and statistical language models allow too many possibilities and thus it can hardly help to reject invalid recognition hypotheses. The original idea of restricting the recognition grammar according to the position in the dialogue scenario was extended to the other parts of the framework and resulted in a general scheme of information retrieval dialogue system with spreading of extracted knowledge.

Acknowledgement

The research presented above was partly supported by a Grant Agency of the Czech Republic under contract No. 201/02/1553 "Natural Language Based Communication with Information and Database Systems". Theoretical part of the work was supported by the Ministry of Education of the Czech Republic under project No. MSM 235200005 – "Intelligent Information Systems and Technologies".

References

[1] Beneš, V.: Sémantická analýza doménově roztříděných dialogů (Semantic Analysis of Domain Dependent Dialogues). M.A. Thesis (in Czech only), University of West Bohemia, Pilsen, Czech Republic, 2003

[2] Ekštein K., Matoušek V., Pavelka T.: Automatische Segmentierung und Markierung des Sprachsignals. In: Tagungsband der Konferenz ESSV 2003, Karlsruhe, September 2003.

[3] Ekštein, K., Mouček, R.: Time-Domain Structural Analysis of Speech. In: Proceedings of CICLing 2003, Mexico City, February 2003.

[4] Ekštein, K., Mouček, R.: Detection of Relevant Speech Features Using Driven Spectral Analysis (the LASER Case). Proceedings (CD) of 4th International PhD. Workshop on Information Technologies and Control. Institute of Information Theory and Automation, Prague, Czech Republic, 2003

[5] Ekštein, K., Mouček, R.: Detection of Relevant Speech Features using Driven Spectral Analysis. In: Proceedings of the 4th International PhD Workshop, Spa Libverda, Czech Republic, September 2003.

[6] Ekštein, K., Hitzenberger, L., Klečková, J., Krutišová, J., Kubišta, J., Matoušek, V., Mouček, R., Taušer, K.: Novel communication concepts for municipal information services. In: SoftCOM 2003: International Conference on Software, Telecommunications and Computer Networks. University of Split, 2003, pp. 705 – 709.

[7] Ekštein K., Pavelka T.: LINGVO/LASER: Prototyping Concept of Dialogue Information System with Spreading Knowledge. In: Proceedings of the Int. Workshop on Natural Language Understanding and Cognitive Science NLUCS 2004, Porto, Portugal, April 2004, pp. 159 – 168.

[8] Lorenzová L.: Psycholingvistická analýza komunikace člověka s dialogovým informačním systémem (Psycholinguistic Analysis of Communication between Human and Dialogue Information System). M.A. Thesis (in Czech only), University of West Bohemia, Pilsen, Czech Republic, 1999.

[9] Matoušek, V., Ocelíková, J.: Managing Spoken Dialogs in Information Services. In: Proceedings of the 7th IFIP TC13 Conference INTERACT '99 on Human-Computer Interaction, Edinburgh, Scotland, September 1999, S. 141-148

[10] Matoušek V., Nöth E.: Ein mehrsprachiges multifunktionelles Auskunftsdialogsystem. In: Mehnert, D.: „Elektronische Sprachsignalverarbeitung“, v.e.b. Universitätsverlag Dresden, September 1999, S. 136-143

[11] Mouček, R., Ekštein, K.: Corpus Construction within Linguistic Module of City Information Dialogue System. In: Proceedings of CICLing 2003, Mexico City, February 2003.

[12] Mouček, R. Ekštein, K.: Municipal Information System. Proceedings (CD) of 4th International PhD. Workshop on Information Technologies and Control. Institute of Information Theory and Automation, Prague, Czech Republic, 2003

[13] Pavelka, T.: Implementation of Hybrid Speech Recognizer. M.A. Thesis (in Czech only), University of West Bohemia, Pilsen, Czech Republic, 2003

[14] Schukat-Talamazzini, E. G., Hendrych, R., Kompe, R., Niemann, H.: Permugram Language Models. In: Proc. of Int. Conference on Artificial Intelligence, Munich, Germany, 1995, pp. 283 – 291.

[15] Schwarz J., Matoušek V.: Automatic Analysis of Real Dialogues and Generation of Training Corpora. In: Proceedings of the Int. Conference EUROSPEECH 2001, Aalborg, Danmark, September 2001, Volume 4, pp. 2201 – 2204.

[16] Speech Experts GmbH: Whitepaper, `www.speechexperts.com`, Regensburg, Germany, 2003