From the Proc. of the *European Congress on Intelligent Techniques and Soft Computing*, EUFIT'95, pp. 1633–1637

TOWARDS DOMAIN–INDEPENDENT UNDERSTANDING OF SPONTANEOUS SPEECH

R. Kompe, W. Eckert, A. Kießling, T. Kuhn, M. Mast,
H. Niemann, E. Nöth, E.G. Schukat–Talamazzini, S. Rieck

Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg
Martensstr. 3, 91058 Erlangen, Germany
Phone: +49 9131 857890, Fax: +49 9131 303811
E–mail: kompe@informatik.uni–erlangen.de

Abstract

We first describe EVAR, a state–of–the–art speech understanding and dialog system which application domain is train time table information retrieval. The system — as others — is devided into two main components: a speech recognizer based on hidden Markov models (a statistical approach) and a knowledge-based linguistic component developed in a semantic network environment. When switching to a new domain the knowledge–base would have to be changed manually, which means a great effort. Thus, we favor for a statistical approach for linguistic analysis. This would only mean retraining the models on annotated corpora from the new domain. A promising approach are semantic classification trees. Since the speech recognizer already relies on statistical models only (part of) its parameters have to be retrained when switching to a new domain.

1 Introduction

Automatic speech recognition (ASR) and automatic speech understanding (ASU) are difficult tasks even for read speech. In ASR one has to deal with acoustic invariances across speakers, coarticulation and sluring. Furthermore, noise, channel characteristica, and the absence of word boundaries in the signal cause problems. In ASU the structure underlying the surface of the word sequence has to be found. This can be done with formal grammars. However, for natural language context–free grammars are not sufficient. Based on this structure underlying the surface, the word sequence has to be interpreted in the specific application domain.

Moreover, in any realistic application, a speech recognition module integrated into a voice dialog system has to cope with the phenomena of *spontaneously* produced speech. Typical phenomena are corrections, repetitions, and false starts; more generally, the degree of grammaticality is expected to be lower for conversational speech than for written texts [11]. Furthermore, we must be prepared to encounter nonverbal speech productions like coughs, lip smacks, and breath noise as well as different kinds of hesitations; words might occur which are outside of the recognition vocabulary. Finally, the pronunciation of spontaneous speech tends to be less careful compared to read speech.

Thus, up to now all known prototypical ASR and ASU systems are limited to a specific application domain. This allows to keep the recognition vocabulary fixed and small, and it allows the use of a restricted linguisitic model. Furthermore, usually the ability to deal with spontaneous speech phenomena is very limited.

In this paper, we first (Section 2) describe the speech understanding and dialog system EVAR, which is a system for train time table inquiries. In Section 3 we give an overview over the problems which occur when switching to a new domain, and we show how these might be overcome in the future.

2 The Speech Understanding and Dialog System EVAR

2.1 Overview

The speech understanding and dialog system EVAR is an experimental automatic travel information system in the domain of German *InterCity* train time table inquiries. Figure 1 shows the structure of EVAR. The two main components are the linguistic analysis and the acoustic processing.

Input to the system is continuous German speech, which is recorded with a DeskLab from Gradient directly connected to the work station where the system is running. No other special hardware is used. In the current version, output of the speech recognition component is the best matching word sequence, but word hypotheses lattices including alternative word hypotheses can be used as well (cf. Section 2.2 and [12]). A knowledge–based approach is used for the linguistic analysis (cf. Section 2.3 and [6]). Its goal is the interpretation of the word

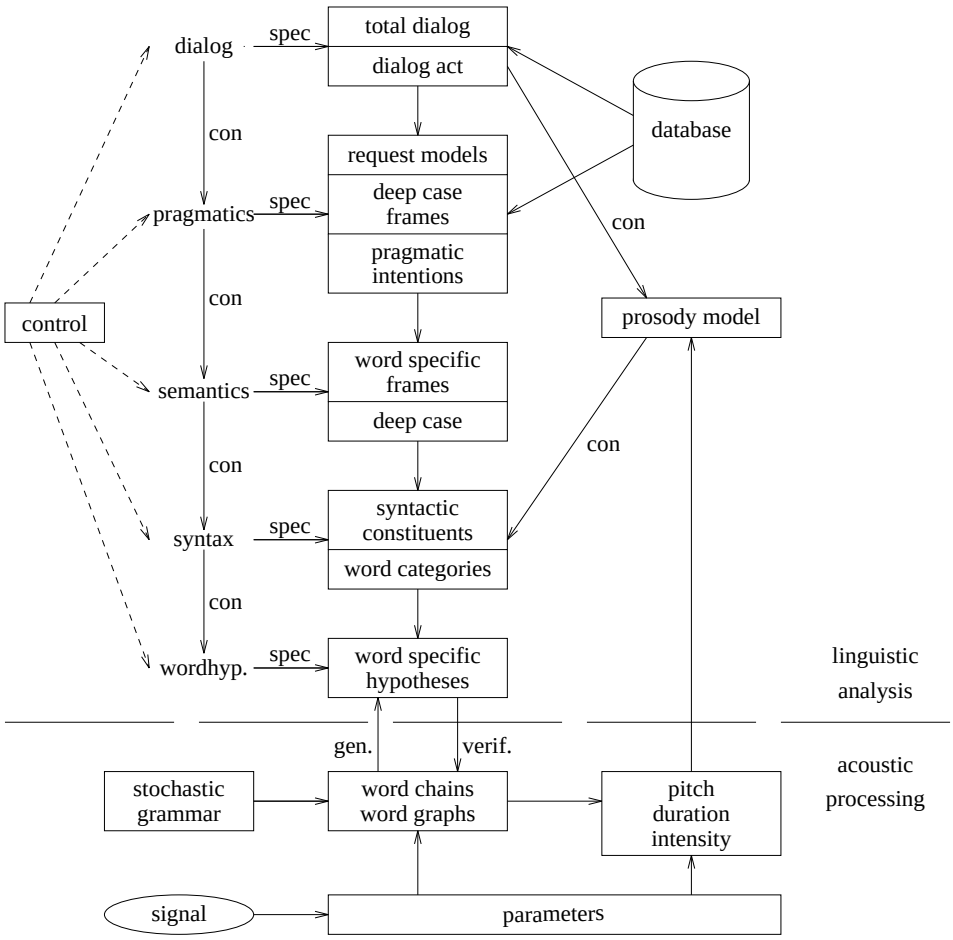

Figure 1: The speech recognition and dialog system EVAR.

sequence computed by the speech recognizer with respect to all the parameters necessary for a database request. This means that not all of the word sequence has to be analyzed. If information is missed a clarification dialog is conducted. Clarifying questions and system answers are given via speech synthesis.

The system uses HaFas, the official database used by the German railway company. This is very important, since the system is connected to the public telephone network in order to test it with naive users. These tests are used to improve the linguistic analysis as well as to acquire more training data for the speech recognizer.

2.2 Recognition of Spontaneous Speech

During feature extraction, a 24–dimensional vector, consisting of the short time log energy, 11 mel–frequency cepstral coefficients as well as its temporal derivatives is computed every 10 msec. In order to cope with the time–varying channel characteristics of different telephone lines, some feature normalization is performed.

The ultimate goal of a statistical speech decoder is to select from all possible word sequences $\boldsymbol{w}$ that sentence hypothesis $\boldsymbol{w}^*$ which maximizes the a posteriori probability $P(\boldsymbol{w}|\boldsymbol{X})$ or, equivalently, the joint probability $P(\boldsymbol{w}, \boldsymbol{X}) = P(\boldsymbol{X}|\boldsymbol{w}) \cdot P(\boldsymbol{w})$ of $\boldsymbol{w}$ and the sequence of acoustic features $\boldsymbol{X}$. The conditional probability function $P(\boldsymbol{X}|\boldsymbol{w})$ is referred to as the *acoustic model* of the decoder; estimates of these quantities are provided in our system by hidden Markov models (HMMs). The a priori sentence probability $P(\boldsymbol{w})$ is referred to as the *language model* of the decoder. The language model is a stochastic linguistic model which represents only the surface structure of sentences. We approximate $P(\boldsymbol{w})$ by a product of smoothed conditional n–gram probabilities (called polygrams): $P(\boldsymbol{w}) = P(w_1 \ldots w_m) = P(w_1) \cdot \prod_{n=1}^{m} P(w_n|w_1 \ldots w_{n-1})$

The search space of a Viterbi speech decoder grows exponentially with the order n of the language model. Thus, we adopted a three-stage recognizer architecture (Figure 2). First, a fast forward Viterbi search transforms the feature vector representation of the speech signal into a word lattice using the acoustic models (HMMs) and a bigram language model. This lattice is searched backward by an A*–algorithm for an ordered list of the N best-scoring sentence hypotheses. Out of these the best word sequence is determined in combining the probabilities of a higher–order language model (polygram) with the ones of the acoustic model.

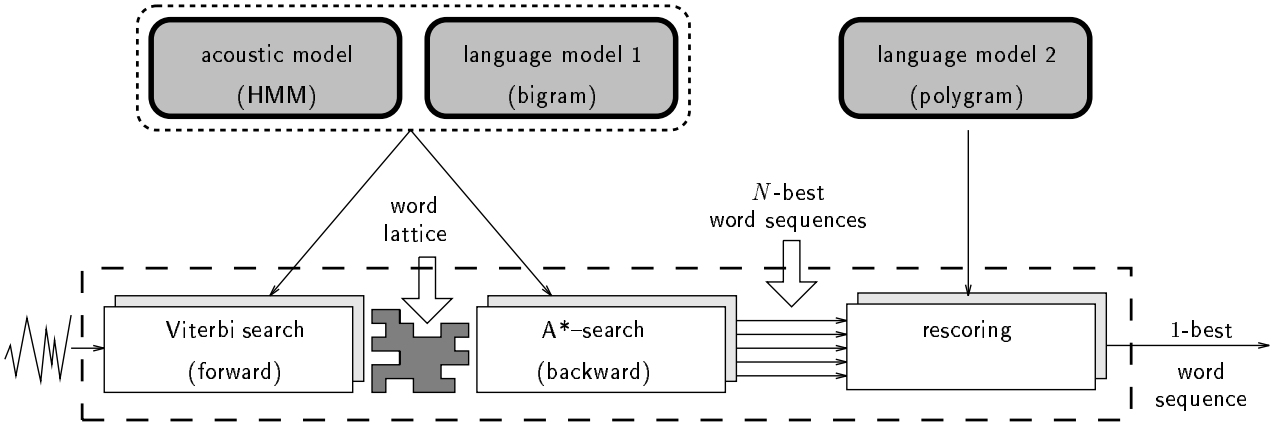

Figure 2: A three–stage architecture for continuous speech recognition.

The acoustic scores rely on HMM for polyphones, which are context–dependent phone models of arbitrary length. The optimal phonetic modelling unit for speech recognition should be small enough to allow robust parameter estimation from a speech database. On the other hand it should be large enough to capture coarticulatory effects sufficiently. Polyphones provide an extremely flexible means for subword unit definition with respect to coarticulation and to portability onto new domains. The optimal set of polyphone units is aquired automatically (3515 different units in the ERBA task, i.e. more than two million statistical parameters).

The statistical parameters of the recognizer were trained on 10100 inquiries read by 100 speakers (14 hours of speech). On speech data from the same domain but from 10 new speakers a word accuracy (WA) of 89.3% and a sentence accuracy (SA) of 59.0% was achieved (cf. column ERBA in Table 1; for the DIPHON results see Section 3.1. For comparison, on real spontaneous speech data obtained from face–to–face dialogs in the appointment scheduling domain of the German VERBMOBIL project a word accuracy of only 45.5% could be achieved (see column VERBMOBIL in Table 1). The training data consisted of the ERBA and the DIPHON data and the VERBMOBIL data not used for testing (alltogether 40 hours of speech). In both domains the vocabulary used for testing consisted of about 2200 words.

2.3 Interpretation of Utterances

For the representation of the knowledge the semantic network system ERNEST is used (cf. [8]). All knowledge needed for the speech understanding process and for the dialog is embedded within a single semantic network using the same representation language. Thus, it is easy to propagate restrictions from all levels to support the recognition process. Nevertheless the knowledge base is easy to extend and to modify because of its modularization into *levels of abstraction*. These levels are connected with *concrete (con)* links. Within a level of abstraction concepts are connected using *part–of* and *specialization (spec)* links. A concept refers to a term or a object. It is defined by attributes and relations. The following modules are integrated (see Figure 1):

The *word hypotheses* module builds up the interface between speech recognition and linguistic analysis.

The *syntactic* module represents special dialog constructions, e.g. "can you tell me", and constituents e.g. the structure of prepositional noun groups (PNG) or time expressions. In German it is characteristic for spontaneous speech that the order of the constituents can be rather free, which is taken into account in the system.

In the *semantic* module, verb and noun frames with their deep cases according to the deep case theory are represented. According to these, the syntactic constituents are interpreted independently from the domain, e.g. time expressions like "a quater to eight" or "the day after tomorrow early in the morning" are mapped to a specific date and time (interval).

The *pragmatic* module represents task–specific knowledge, i.e. a description of terms like "place of arrival" (see Figure 3) or "to go" (by train). With these the parameters for the database request are extracted. Also for prediction purposes the pragmatic knowledge is useful, e.g. if a city name is interpreted as a goal on the semantic level (S_GOAL) and as place of arrival on the pragmatic level then it has to be preceded by the preposition "nach".

The *dialog* module represents all expected sequences of dialog acts. These have been manually derived from a set of human–human dialogs out of the domain of train table inquiry. The dialog memory keeps track of the dialog history, since a user utterance has not to be interpreted in the dialog context, e.g. to resolve anaphoric references and to focus the analysis on the expected answer.

The *prosody* module determines sentence mood, accentuated words and phrase boundaries on the basis

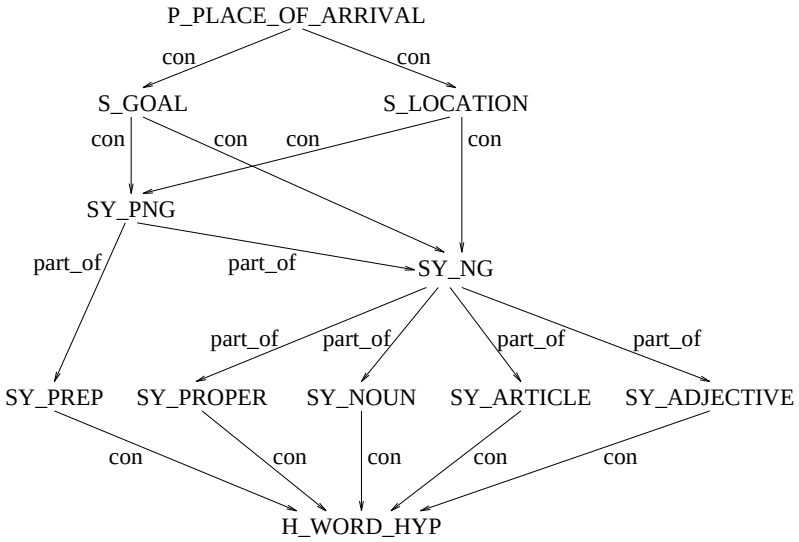

Figure 3: Model of the term "place of arrival" in the semantic network. The concepts of the pragmatic, semantic, syntactic levels are indicated by the prefixes P, S, and SY respectively.

of changes in pitch, duration, and intensity. Currently only sentence mood is used in the linguistic analysis. The user has the opportunity to repeat time–of–day expressions given in the system answer. In these elliptic repetitions only the pitch contour determines the sentence mood, e.g. if the user uttered a question (which means "please confirm, if I understood correctly") [2].

For the *control* of the analysis process, ERNEST provides a problem–independent procedural semantics of the network language, which is flexible so that the search can alternate between bottom–up (interpretation) and top–down (prediction) mode. The A*–algorithm in combination with problem–dependent judgement vectors is the basis of this control. Alternative intermediate interpretations are considered simultanously. In each analysis step the best alternative is expanded. For determining the "best" alternative, the scores computed by the speech recognizer are combined with the linguistic judgements. The ultimate goal of the analysis is the instantiation of a sequence of dialog–level concepts until all the parameters for a database request are known. A concept can be instantiated when one out of a collection of predefined subsets of concepts could be instantiated to which it is connected via *concrete* or *part–of* links.

3 Towards Domain Independency

3.1 Speech Recognition

The results in Table 1 are given for a speech database read by 10 speakers from the train time table information retrieval domain. We trained two speech recognizers, one on the ERBA database from the same domain (14 hours of read speech) and the other on the DIPHON database (20 hours of read speech) from a different domain. When switching to a new domain the recognition performance of word recognizers drops significantly when no language model is available (see column DIPHON). This is due to the fact that the recognition erformance relies very much on representative context dependent phone models; their occurence of is to a great deal domain or — to be more specific — vocabulary dependent. Using a domain dependent bigram or even polygram language model increases the recogntion rates considerably and moreover there is not much difference anymore between the results in the two columns DIPHON and ERBA. Thus, for speech recognition it seems to be sufficient to train a statistical language model on the new domain which needs only little effort. There exist also approaches where language models adapt to new domains starting from a general language model, e.g. the cache–based language model [3].

For state–of–the–art speech recognizers a pronunciation lexicon is required for all the words to be recognized. Usually these lexica are built manually. How to deal with unknown words is still an open question and an important research topic. Even the problem of recognizing which segments of speech belong to unknown words is a non–trivial task. Furthermore, an open problem is how to adapt automatically to new domains.

3.2 Semantic Interpretation

So far, in most prototype systems known to the authors, the speech understanding component is knowledge based as it is the case in EVAR. A drawback of these systems is that the knowledge is manually acquired and

language model	DIPHON		ERBA		VERBMOBIL
	WA	SA	WA	SA	WA
none	40.0	3.2	58.5	7.6	—
bigram	81.7	43.4	83.6	37.4	—
polygram	87.4	58.4	89.3	59.0	45.4

Table 1: Word (WA) and sentence (SA) accuracies of our speech recognizers using different parameter sets. Columns DIPHON and ERBA refer to read speech, the same testing data but different trained data has been used. VERBMOBIL refers to a task, where training and testing data were spontaneous speech data. Accuracy is defined as % correct – % inserted – % deleted.

highly domain dependent. This makes switching to a new domain extremely time consuming. Thus, several research groups started activities towards statistical models for seamtic interpretation, which can be trained automatically on the basis of annotated corpora. The most known approaches are probabilistic context–free grammars (cf. e.g. [1]), Pieraccini's approach [10], and semantic classification trees (SCT) [4].

SCTs are decision trees originally developed for the semantic classification of word sequences. The nodes of the trees are associated with binary questions (rules). The questions are about subsets of word strings represented as regular expressions. The specific questions and the order in which they are applied are determined during training. The leaves of the trees correspond to the semantic classes with a certain probability. SCTs have been successfully used by R. Kuhn in a speech understanding system in the domain of flight information retrieval [4]. We used SCTs for the detection of syntactic phrase boundaries in word sequences. For three classes of boundaries a recognition rate of 92% could be achieved on a test set different from the training set. We also started using SCTs for the classification of dialog acts [7]. In all of these tasks SCTs showed very good generalization performance.

4 Conclusion

Speech understanding and dialog systems comparable to EVAR have also been built by other groups. Meanwhile the performance of most of them is well enough to build research prototypes which can be used by naive but cooperative users. Using them in real applications seems not anymore to be far in the future. There might be a lot of useful application domains, especially all kinds of database inquiries, speech–to–speech translation, customer advice service, interfaces to intelligent robots [5], etc. At present it means a lot of effort to switch from one of these domains to the other, because most of the linguistic components working reasonably well are based on manually acquired knowledge. Thus, a major research effort in the field of speech understanding concerns the automatic aquisition of knowlegde. The first approaches towards this goal have been proposed by several research groups; all of them rely on statistical methods.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research projects ASL and VERBMOBIL, by the German Research Foundation, and by the European Comission in the ESPRIT/SUNDIAL project. Only the authors are responsible for the contents.

References

[1] A. Corraza, R. de Mori, R. Gretter, G. Satta. Recent results on stochastic language modelling. In [9], pages 45–52.
[2] R. Kompe, E. Nöth, A. Kießling, T. Kuhn, M. Mast, H. Niemann, K. Ott, A. Batliner. Prosody takes over: Towards a prosodically guided dialog system. *Speech Communication*, 15(1-2):155–167, Oktober 1994.
[3] R. Kuhn, R. De Mori. A Cache-Based Natural Language Model for Speech Recognition. *IEEE Trans. PAMI*, 570–583, 1990.
[4] R. Kuhn, R. de Mori. Learning Speech Semantics with String Classification Trees. In *Int. Conf. on Spoken Language Processing*, volume 2, pages 55–58, Minneapolis, April 1993.
[5] G. Lazzari. Automatic speech recognition and understanding at IRST. In [9], pages 149–157.
[6] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, G. Sagerer. A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base. *IEEE Trans. PAMI*, 179–16, 1994.
[7] M. Mast, E. Nöth, H. Niemann, E.G. Schukat-Talamazzini. Automatic Classification of Speech Acts with Semantic Classification Trees and Polygrams. In *IJCAI-95 Workshop "New Approaches to Learning for Natural Language Processing"*.
[8] H. Niemann, G. Sagerer, S. Schröder, F. Kummert. ERNEST: A Semantic Network System for Pattern Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 883–905, 1990.
[9] Niemann, de Mori, Hanrieder, editors, *Progress and Prospects of Speech Research and Technology: Proc. of the CRIM/FORWISS Workshop (München, Sept. 1994)*, pages 45–52, Sankt Augustin, 1994. infix.
[10] R. Pierrachini. Speech understanding and dialog, a stochastic approach. In A. Rubio, editor, *New Advances and Trends in Speech Recognition and Coding*, NATO ASI Series F. Springer–Verlag, Berlin, Heidelberg, New York, 1995.
[11] S. Schachtl, H.-U. Block. Syntaktische Beschreibung in Systemen zur Verarbeitung gesprochener Sprache. Technical report, ASL–TR–11–91/SIM, September 1991.
[12] E.G. Schukat-Talamazzini, T. Kuhn, H. Niemann. Speech Recognition for Spoken Dialogue Systems. In [9], pages 110–120.