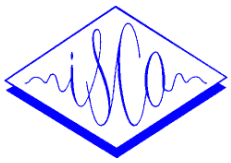

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

SPICOS II - A SPEECH UNDERSTANDING DIALOGUE SYSTEM

Harald Höge

Siemens AG
D-8000 Munich, Germany

ABSTRACT

SPICOS II represents a speaker adaptive speech understanding dialogue system allowing data base requests. The system consists of a dialogue manager, components for acoustic and linguistic analysis and components for data base access and answer generation. The acoustic analysis is based on a speaker adaptive articulatory feature vector and speaker-independent HMM-phoneme models. The linguistic analysis is carried out using an augmented phrase structure grammar and a formal-logic semantic representation.

I. INTRODUCTION

In 1984 the companies Siemens and Philips started a joint project SPICOS (Siemens-Philips-IPO Continuous Speech Recognition and Understanding) integrating the resources in speech recognition and computer linguistic. In the meantime two prototypes SPICOS I [1] and SPICOS II of speech understanding systems have been developed which allow the retrieval of the contents of a database in fluently spoken natural speech. The main features of the two systems are shown in Tab. 1.
This paper describes SPICOS II. Concerning the acoustic and linguistic scope SPICOS II can be compared with the Voyager System [2]. The following dialogue sequence gives an impression of the capabilities of SPICOS II.

1 U(ser): hat Herr Doktor Höge den Brief über Akustik an Herrn Noll geschickt
(did Doctor Höge send the letter on acoustics to Mr. Noll)

1 S(ystem): haben Sie gesagt: hat Herr Doktor Höge den Brief über Akustik an Herrn Noll geschickt
(did you say: did Doctor Höge send the letter on acoustics to Mr. Noll)

2 U: ja
(yes)

2 S: ja Höge hat den Brief geschickt
(yes Höge sent the letter)

3 U: war er auf dem letzten Treffen in Hamburg
(was he on the last meeting in Hamburg)

3 S: meinen Sie mit 'er' Höge oder Noll
(by 'he' do you mean Höge or Noll)

4 U: Höge
(Höge)

4 S: nein Höge war nicht auf dem letzen Treffen
(no Höge was not at the last meeting)

System Properties	*SPICOS I*	*SPICOS II*
Task	database inquiries	
Language	German	
Vocabulary Size (full forms)	917	1150
Speech Input	continuous, 6.4kHz bandwidth	
	speaker	speaker dependent/ speaker adaptive
Acoustic Analysis	discrete HMM	continuous HMM
	phoneme models	
Language Model	network grammar	statistical bigram grammar
Search	top 1 search	Top-N search fast look ahead
Dialog structure	one question one answer	clarification dialogue follow-up questions
Linguistic Analysis	augmented phrase structure grammar formal semantic representation	
		presupposition checking interactive disambiguation
Speech Synthesis	male voice	greater intelligebility female voice
System Response	100 times real time	20 times real time
Prototype	1986	1990

Tab. 1 - Features of SPICOS I and SPICOS II Systems

Besides information requests and supplies the above example contains stages involving clarification and verification (1S, 2U, 3S, 4U). The latter are the result of uncertainties in the system (1S) and from ambiguously resolvable situations like (3S). The permissible user reactions are either simple yes/no-answers to the systems questions or one-word replies (4U). If the recognition is uncertain an echo-question (1S) is put by the system in order to obtain verification. (3S) is induced by a genuine ambiguity in the anaphora resolution mode.

The architecture of SPICOS II is shown in Fig. 1.

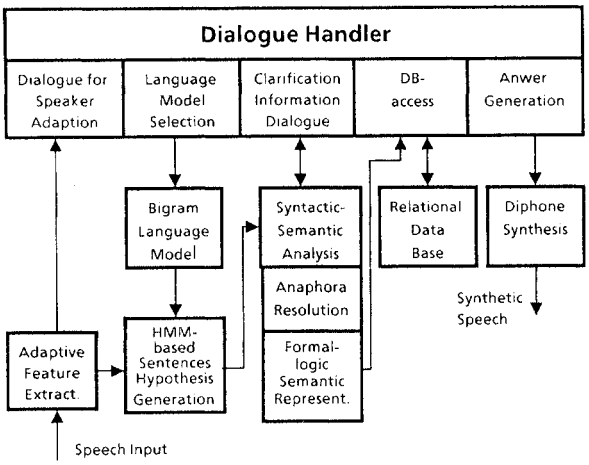

Fig. 1 - Architecture of SPICOS II

The control of the system is organized by a dialogue handler, which is implemented as a finite state machine [3]. In each state specific dialogue procedures are performed. In the "speaker adaptation state" a new speaker is asked to speak two selected sentences with which the system is adapted [4]. If the dialogue handler starts the recognition module, a dialogue state specific language model is activated. In the case of an "information dialogue state" a language model, which covers the whole DB-domain, is selected [5]. In a verification or clarification dialogue state a very simple language model, allowing only single word replies is valid. The recognition itself is based on context independent HMM phoneme models [6] and a top N-beam search strategy [7] augmented with a fast look ahead procedure [8]. In the actual implementation the recognition module delivers 4 scored sentence hypotheses for each uttered phrase. Out of these hypotheses the linguistic module selects the highest scored sentence which is compatible with the systems syntactic and semantic coverage [9]. If no hypothesis is compatible the dialogue handler asks the user to repeat the question. During the "information dialogue state" the main task of syntactic and semantic analysis is the mapping of a user question to a data base request. A formal-logic semantic representation is used as data base query language. Further a discourse representation is build up during the whole man-machine dialogue. This discourse representation is needed for the resolution of anaphoric expressions. Syntactic-semantic ambigue situations are resolved during the "clarification dialogue state". If a query is answered, the answer pattern is generated according to the type of input question (y/n-question, wh-question, etc.), the result of the database evaluation and the type of presupposition failures is contained in the sentence. The answer pattern is produced by an appropriate transformation of the structure of the input question [10]. The speech synthesis is performed by a diphone synthesis developed at IPO [11].

II. SPEAKER ADAPTIVE SPEECH RECOGNITION

Because the difference in recognition rate between speaker dependent and speaker independent recognition systems is significant, speaker adaptive recognition procedures are an attractive alternative. Most speaker adaptive systems transform speaker specific spectral parameters into a normalized feature space, where the transformation has to be learned by a few utterances. These normalized features are input to a conventional HMM recognizer.
Our main idea was to use articulatory features as normalized features. In our system the articulatory process is described by an articulatory state defined by the categories of manner and place of articulation (see Tab. 2/3).

Consonants				
LA	=	labial	:	/b/, /p/, /m/
LD	=	labio-dental	:	/v/, /f/
DA	=	dental	:	/z/, /s/, /t/
AL	=	alveolar	:	/d/, /l/, /n/
PA	=	palatal	:	/sch/, /ch/, /rj/, /j/
VE	=	velar	:	/ng/, /x/, /rx/, /g/, /k/
GL	=	glottal	:	/?/, /h/
Vowels and Diphtongs				
		F1	**F2**	
V1	=	low,	low :	/u/, /uh/, /o/, /oh/
V2	=	high,	low :	/a/, /ah/, /er/
V3	=	central,	central:	/9/, /ae/, /oe/, /ue/
V4	=	low,	central:	/ueh/, /oeh/
V5	=	low,	high :	/i/, /ih/, /eh/

Tab. 2 - Categories of place of articulation

SI	=	pause, closure	:	/si/, /t/, /p/, /k/
WF	=	weak frikativ	:	/?/, /h/, /v/, /z/, /t/, /rj/, /rx/, /p/, /k/
SF	=	strong frikativ	:	/sch/, /ch/, /x/, /s/, /f/
VB	=	voice bar	:	/b/, /d/, /g/
NA	=	nasal	:	/m/, /n/, /ng/
SO	=	sonorant	:	/l/, /j/, /r/,
VO	=	vowel	:	vowels and diphtongs

Tab. 3 - Categories of manner of articulation

The articulatory categories are per definition speaker independent, are closely related to the phonemes and their number is small. These properties have been the main reason to use articulatory categories as normalized features.

As it is well known their exist no reliable method to determine the articulatory state from the speech signal directly, although much progress has been made in the last years [12]. For this reason we choose a probabilistic approach which states that only the probability for an articulatory state for a given speech segment can be calculated. For example the articulatory state (manner of articulation = AL, place of articulation = VB), which corresponds to the phoneme d, has a certain probability P (articulatory state = AL, VB). We assume that manner and place of articulation are statistic independent processes leading to the general expression P (articulatory state) = P (manner of articulation) P (place of articulation). Within this scheme the probabilities of articulatory states can be calculated from the 19 dimensional probabilistic feature vector AFV = (P(LA), P (LD), . . . , P (VO)), which is the definition of our articulatory feature vector. The mapping of the speech signal to the AFV vector is treated as a classification problem [4]. The probabilities for manner of articulation are estimated from robust signal parameters as energy contour, zero crossing and broad spectral energy ratios. The probabilities of place of articulation are derived from the first 4 formants. The classifying itself is done with a HMM procedure, where each category is modeled by 3 states described by gaussian mixture densities.
In this context speaker adaption is equivalent with the adjustment of these gaussian densities to a new speaker, i.e. mean values and variancies have to be readjusted. From the nature of the signal parameters chosen two sentences are sufficient to make a proper transformation. For comparison neural net classifier have been investigated [13]. But this approach needs a longer adaption time and labeled data.
The articulatory vector has been easily integrated in the HMM phoneme model [6], leading to an recognition scheme for top-N sentence hypotheses.

III. LINGUISTIC PROCESSING

Besides answer generation linguistic processing performs syntactic and semantic analysis providing data base access, anaphoric resolution, presupposition handling and rejection of wrong sentence hypotheses [9].
A modified left to right top down chart parser handling discontinuous phrase structures is used for syntactic analysis. The parser works with an feature augmented phrase structure grammar APSG, which is supported by a semantic network checking the semantic compatibility of content words. This semantically enriched syntactic analysis is devoted on the one hand side to reject wrong sentence hypotheses and on the other side to build up a syntactic tree for further semantic processing and discourse analysis. The current syntax covers imperative and declarative sentences, yes/no- and wh-questions, main and relative clauses, active voice and present, past, perfect. The grammar comprises 100 main rules with 175 subrules, 12 terminal and 27 non terminal categories and a comprehensive feature system. Besides checking the semantic compatibility of words the semantic network performs deep case analysis used for the formal semantic representation and attaches semantic categories to semantically empty pronouns for supporting anaphora resolution.
The anaphoric resolution module collects nominal phrases as antecedent-candidates and anaphoric expressions. The linking of an anaphoric expression to its appropriate antecedent is performed by a reference grammar.
The semantic analysis works as a two stage process. The first stage aims at a representation of meaning, which is independent from the domain of discourse. Basically this means that the words of the language function as atoms of semantic representation. This formal representation contains the ambiguity and vagueness of the natural language. In a second step this vague semantic representation is transformed to a domain specific precise semantic representation, where the atoms are converted to the special structures of the underlying data base model. These two representations are variants of the ELF/ELR (Ensemble Language Formal/Ensemble Language referential representations which are extensively described in Bunt [15].

IV. THE SPICOS II DEMONSTRATOR

SPICOS II is implemented on 3 workstations with dedicated hardware [14] (see Fig.2).

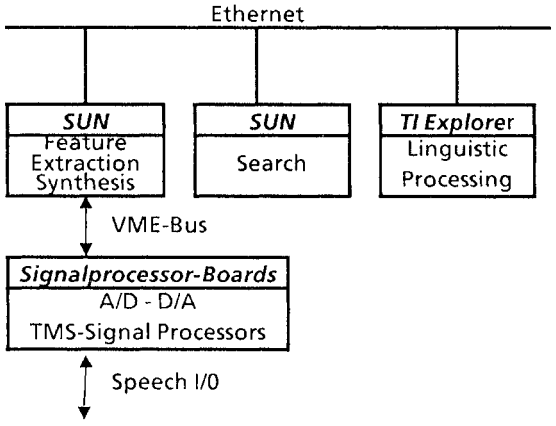

Fig. 2 - Hardware Architecture of SPICOS II

The distributed hardware architecture was chosen to come closer to real time performance which needs in the current implementation about 230 MIPS. With the hardware shown in Fig. 2 the analysis of 10 seconds of speech needs about 200 seconds. The main processing power is consumed for the N-top beam search.
SPICOS II is now in its evaluation phase. For testing and training a SPICOS II speech data base for 12 speakers has been created. For each speaker 2 sessions of 100 phonetic balanced sentences (training corpus) and 50 domain specific dialogue sequences (test corpus) have been recorded. For the used language model a test set perplexity of 220 with a resulting sentence recognition rate (top 1) of 26 % to 51 % was measured.

V. CONCLUSION

SPICOS II is a step toward man-machine interfaces for data base inquiries. Yet much work has to be done to improve recognition rate, linguistic coverage and real time performance.
This research was supported by the German BMFT under contract ITM 8401. The author alone is responsible for the content.

REFERENCES

[1] H.Höge, H.Ney: "Architektur des sprachverstehenden Systems SPICOS", Proc. Kleinheubacher Berichte, pp. 29-36, 1985.

[2] V.Zue et.al.: "The Voyager Speech Understanding System: Preliminary Development and Evaluation", Proc. ICASSP-90, pp. 73-76,1990.

[3] J.de Vet, K.van Deemter: "A Dialogue Handler for SPICOS II", IPO-Report, Eindhoven, 1988.

[4] O.Schmidbauer: "Robust statistic modelling of systematic variabilities in continuous speech incorporation acoustic-articulatory relations", Proc. ICASSP '89, Glasgow, 42 S. 12.5, 1989.

[5] A.Paeseler, H.Ney: "Continuous speech recognition using a stochastic language model", Proc. ICASSP '89, Glasgow, 44 S. 13.7, 1989.

[6] H.Ney, A.Noll: "Phoneme Modelling using continuous mixture densities", Proc. ICASSP '88, pp.437-440, 1988.

[7] V.Steinbiss: "Sentence-Hypotheses Generation in a continuous Speech Recognition System", Proc. Eurospeech '89, Vol. 2, pp.51-54, 1989.

[8] X.L.Aubert: "Fast Look-Ahead Prunning Strategies in Continuous Speech Recognition", Proc. ICASSP '89, pp. 659-662, 1989.

[9] G.Niedermair, M.Streit, H.Tropf: "Linguistic Processing Related to speech Understanding in SPICOS II" to be published ed. M. Wajskop, special issue of Speech Communication Journal, 1990.

[10] J.de Vet, et.al.: "The SPICOS II Answer Generator", IPO-Report 703, 1988.

[11] J.Van Hemert, et.al.: "Speech Synthesis in the SPICOS project", eds. H.G.Tillmann, Wille: Analyse u. Synthese gesprochener Sprache. Linguistische Datenverarbeitung, Vol. 9, pp.34-39, 1987.

[12] J.Schroeter, et.al.: "Multi-Frame Approach for Parameter Estimation of a Physiological Model of Speech Production", Proc. ICASSP '88, pp.83-86, 1988.

[13] A.Aktas, et.al.: "Classification of Coarse Phonetic Categories in Continuous Speech: Statistical Classifiers vs. Temporal Flow Connectionist Network", Proc. ICASSP '90, pp. 89-92, 1990.

[14] A.Aktas, H.Höge: "Multi-DSP and VQ-ASIC Based Acoustic Front-End for Real-Time Speech Processing Tasks", Proc. Eurospeech '89, pp.586-589, 1989.

[15] H.Bunt: "Mass terms and model-theoretic semantics", Cambridge University Press, Cambridge, 1985.