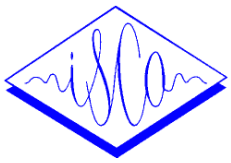

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Robust Speech Understanding

Astrid Brietzmann, Fritz Class, Ute Ehrlich, Paul Heisterkamp
Alfred Kaltenmeier, Klaus Mecklenburg, Peter Regel-Brietzmann
Daimler Benz AG, Research Institute, 89075 Ulm, Germany
name@dbag.ulm.DaimlerBenz.COM

Gerhard Hanrieder, Waltraud Hiltl
FORWISS, 91058 Erlangen, Germany
name@forwiss.uni-erlangen.de

ABSTRACT

This paper describes a fully operational system for analyzing all aspects of continuous speech from word recognition up to linguistic representation. Many systems rely on fully grammatical speech input, others use only shallow analysis without using declarative linguistic knowledge. We propose a flexible processing, where the depth of the analysis varies according to internal (quality of the speech input) and/or external (limitations, e.g. of processing time) criteria. In architecture and system design, special effort was made to cover effects of spontaneous speech (e.g. unknown words, pauses, or sentence breaks). The described system is part of a speech dialog system not discussed in this paper.

1 Requirements for a Robust Speech Understanding System

Although most users of speech dialog systems are cooperative, they are not always experienced in conversing with an automatic dialog system. Thus in real world applications we have to deal with effects like open vocabulary, sentence breaks and ungrammatical sentences. In addition, a recognition system should be able to deal with different speakers over telephone lines. These requirements make *speech analysis* very difficult. Therefore, word recognition and linguistic analysis must be very closely coordinated and must make use of all available restrictions from higher level modules.

2 System Modules

The system described in this paper is part of a dialog system consisting of the following components: speech analysis including linguistic analysis, context-dependent interpretation, dialog strategy, a task module, answer generation, and speech synthesis. The speech analysis produces a linguistic surface structure and a deep structure without regard to the context. The context-dependent interpretation, the dialog and the task modules are based on the ESPRIT Project SUNDIAL [1]. The demonstration system is implemented on a SUN Sparcstation 10. For the domain of InterCity train time table queries (with about 800 word forms) the system requires about six times the utterance time.

3 Control and Search

We propose a flexible hierarchical structure that achieves the best overall results for a given utterance. Depending on the speech input, the analysis level varies from full linguistic analysis, via

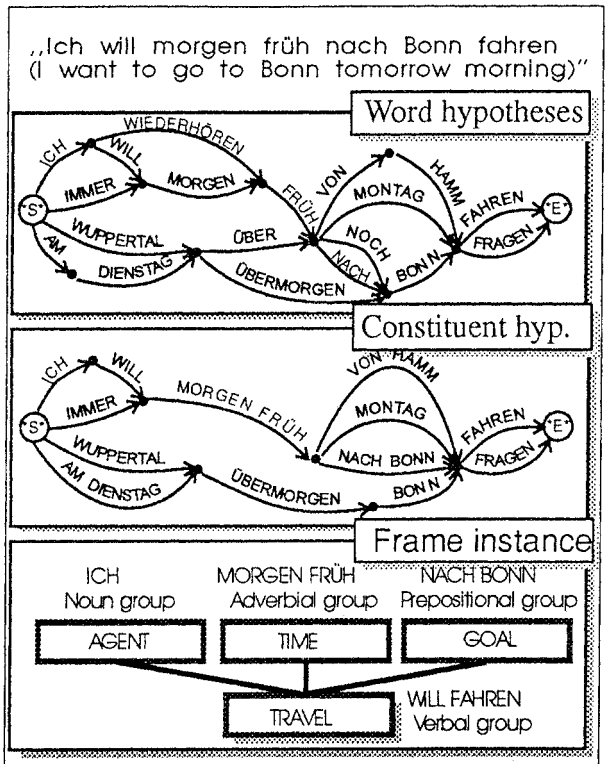

Fig. 1: Steps of the linguistic analysis

analysis of partial sentences (structure spotting) to keyword spotting. The level of analysis is either predetermined (e.g. from the dialog) or is a result of the analysis process. An incomplete analysis may be caused by a sentence construction that does not fit the grammar, by usage of unknown words, or by a huge search space together with a limit on processing time. The analysis begins with keywords, then builds structures around the keywords and as long as no timeout is called by the control module, whole sentences are constructed, if possible. Since fully integrated search is not feasable, search is performed in multiple stages. The acoustic front end uses a 10 ms beam search algorithm to obtain a word hypotheses graph [2]. In this graph, scored hypotheses (lexical words, unknown words, and pauses) are represented as arcs (cf. Fig. 1). In the following stages, keyword and structure spotting techniques together with an *admissable* search strategy controlled by linguistic knowledge find interpretable utterance parts — up to whole utterance hypotheses. The interface between linguistic analysis and domain-dependent interpretation consists of multiple underspecified semantic structures.

4 Acoustic Front End

The recognizer is based on a linear-discriminant transform, soft-decision vector quantization, and semi-continuous Hidden Markov Models (SCHMM) of context-dependent subword units.

4.1 Feature Extraction, Transformation, and Classification

Feature extraction generates 12 mel-based cepstral coefficients and one continuously companded log-energy value every 10 milliseconds. Cepstral features are then normalized using high-pass filters with variable time constants in order to remove the long-time average speech spectrum and spectral influences of the microphone, room acoustics, and transmission line [3]. After normalization, feature vectors are transformed by a linear transform generated by Linear Discriminant Analysis (LDA) [4]. The basic (and new) idea of our approach is to combine the feature vectors of adjacent frames into a new feature vector which implicitely includes temporal dynamic effects of the speech signal. For LDA, 9 frame vectors form a $9 \times 13 = 117$-dimensional vector which is transformed into 32 coefficients by the [32×117] transform matrix.

LDA requires class-specific covariance matrices. In our case, classes correspond directly to SCHMM states or state clusters of 246 subword units. State-dependent covariance matrices are computed during the last iteration of a first forward/backward training based, as in other SCHMM systems, on 4 independent soft-decision vector quantizers (codebooks). However, as opposed to other systems, the first training pass in our system only yields the LDA transform matrix and one codebook for transformed, 32-dimensional vectors. HMM adaptation is then repeated in a second pass using the transformed codebook. The adaptation procedure is described in detail in [3].

4.2 Word Recognition

The recognition vocabulary (word lexicon) is represented as a tree structure of HMM subword units. Each node of the tree represents a subword unit. The recognizer is based on a word-dependent N-best Viterbi algorithm [5]; a bigram language model can be optionally integrated. The output is a word graph whose nodes correspond to time frames and whose arcs are word hypotheses associated with acoustic scores. Likewise word recognition errors, i.e. the recognizer performance, are computed according to the word graph. They are simply defined as the minimum Levenshtein distance between the spoken word string and the word strings along all possible paths through the word graph. Thus the error rate counts insertions, deletions, and substitutions of words. Results are given in [8].

The word graph is updated *frame by frame*. A new graph node is inserted at word ends if the longest path to a new word end is longer than all already existing paths. The word graph implicitely contains a large number of sentences from which the linguistic postprocessor can search the best one with respect to syntactic, semantic, and pragmatic constraints.

5 Non-word Models

Continuous speech recognition must deal with out-of-vocabulary words, hesitiations, and sentence breaks. These effects are accounted for by some so-called non-word or garbage models. In our case, three non-word models with 10, 15, and 25 states are used which model short, medium-length or longer words, respectively. Non-word models are trained with a large number of words alloted to one of the three model classes.

6 Linguistic Analysis

Linguistic analysis fulfills two very important tasks within the system: On the one hand, it provides top down restrictions for the word recognition stage and reduces input ambiguity on the word level. On the other hand, it discovers the most plausible linguistic interpretation of the input for further dialog and task-oriented processing. Considering the huge search space needed for robust continuous word recognition, the search in the word graph hypotheses must be performed within one integrated process.

The linguistic processing component consists of the linguistic knowledge base (lexicon and grammar) and the word graph parser. The linguistic knowledge comprises both syntactic and semantic aspects. The parser combines the island parsing technique ([6]) with a UCG-Parser ([7]) based on the theory of Unification Categorial Grammar. It receives as an input a graph of scored word hypotheses (see section 3). The graph may contain pause-markers and markers for words out of the vocabulary. Non-speech sounds (excessive noise, paralinguistic sounds) are mapped as pauses ([8]).

UCG combines insights of Categorial Grammar with unification-based grammar approaches [9]. Being a categorial grammar, the amount of grammar rules is restricted to a few basic rules of combinations. Most of the combinatorics of words is encoded in the lexical categories. Being a unification grammar, lexical entries are represented as complex feature structures, which are combined by simple unification.

The module is implemented as a constrained based, island driven chart parser using the syntactic and semantic constraints encoded in the UCG grammar. The initial edges are built from the word hypotheses. Middle-out processing enables the system to start from the most promising hypotheses — i.e. high scored word hypotheses matching to top-down predictions provided by the dialog module — and cope with unrecognizable input sections. Since the chart is a monotonic structure, all partial interpretations remain available if the search for a complete utterance hypothesis fails. The analysis is guided by heuristic search with a scoring function. For experimental purposes, several search strategies and scoring schemes have been tried. We achieved good results with a strategy corresponding to the A*-algorithm [10]. The scoring function is a modified version of the *Shortfall*

Score introduced by W. Woods [11].

The parsing is finished when a full description is found, a full search has been done, or a time-out signal is received (pre-set or dialog dependent). Due to recognition errors and/or spontaneous speech phenomena the parser cannot find a complete hypothesis in any case, i.e. all sentence hypotheses are ungrammatical. Instead of simply failing, a **robust** parser delivers partial solutions, which are in many cases sufficient for understanding the utterance [12]. For the selection of partial solutions which should be accepted as a parse result a heuristic quality score is assigned to each chart edge during parsing. This score combines acoustic and linguistic quality measures:

- the shortfall score of the partial result
- the length of edge
- the syntactic completeness and
- the pragmatic relevance of the edge

After parsing, results are generated as follows: Starting from the edge with the best quality score, the best scored left and right adjacent edges are collected recursively until the start and the end of the graph is reached. The results are passed to the dialogue level. Multiple solutions (possibly incomplete semantic structures based on partial analysis results) are delivered to higher-level modules, since the decision for the pragmatically adequate utterance hypothesis cannot be made on purely syntactic and semantic grounds,

The output is further processed by a dialog management module that can continue a reasonable dialog on the basis of incomplete input ([13]).

References

[1] Heisterkamp, P.: Ambiguity and Uncertainty in Spoken Dialog. In: *[14].*

[2] Kaltenmeier, A.: Modellbasierte Worterkennung in Spracherkennungssystemen für großen Wortschatz. Düsseldorf: VDI Verlag 1991.

[3] Class, F.; Kaltenmeier, A.; Regel, P.: Optimization of an HMM-based continuous Speech Recognizer. In: *[14].*

[4] Fukunaga, K.: Introduction to Statistical Pattern Recognition. New York: Academic Pr. 1990.

[5] Schwartz R.; Austin, S.: A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses. In: *Proceedings ICASSP 91*, Toronto 1991.

[6] Brietzmann, A.: Reif für die Insel: Syntaktische Analyse natürlich gesprochener Sprache durch bidirektionales Chart-Parsing. In: *Mangold, H. (ed.)*: Sprachliche Mensch-Maschine-Kommunikation, München, Wien: Oldenbourg 1992.

[7] Andry, F.; Thornton, S.: A parser for speech lattices using a UCG grammar. In: *Proceedings of EUROSPEECH '91*, Genoa 1991.

[8] Class, F.;Kaltenmeier, A.; Regel, P.: Evaluation of an HMM Speech Recognizer with Various Continuous Speech Databases. In: *[14].*

[9] Shieber, S. An Introduction To Unification-based Approaches To Grammar. Stanford: CSLI. (CSLI Lecture Notes. 4.), 1986.

[10] Kanal L.; Kumar, V. (eds.): Search in Artificial Intelligence. New York: Springer 1988.

[11] Woods, W. A.: Optimal Search Strategies for Speech Understanding Control. In: *Artificial Intelligence 18*, 1982.

[12] Baggia, P.; Rullent, C. Partial parsing as a robust parsing strategy. In *Proceedings of ICASSP 93*, Minneapolis, p. 123–126, 1993.

[13] Hanrieder, G.; Heisterkamp, P.: Robust Analysis and Interpretation in Speech Dialogue. In: *Proceedings of CRIM/FORWISS Workshop 1994*, Munich, to appear.

[14] Proceedings of EUROSPEECH '93. Berlin 1993.