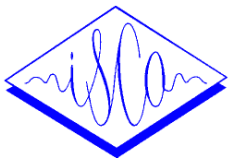

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A CLOSE HIGH-LEVEL INTERACTION SCHEME FOR RECOGNITION AND INTERPRETATION OF SPEECH

Gernot A. Fink, Franz Kummert, Gerhard Sagerer

Universität Bielefeld, Technische Fakultät, AG Angewandte Informatik
Postfach 100 131, 33501 Bielefeld, Germany
Tel.: +49 521 106 2935, Fax: +49 521 106 2992

ABSTRACT

The vast majority of speech understanding systems suffers from a bottleneck between the recognition and the interpretation components. Normally, only a relatively small set of word hypotheses is passed from the recognizer and *no* flow of information in the opposite direction is even possible.

We propose an interaction scheme that tries to overcome many of the disadvantages of traditional systems. It makes use of the possibility to process abstract constituents in our word recognizer and pass them back as complex hypotheses. Predictions that define the complex analysis goal of the recognition can be derived dynamically *during* the interpretation of an utterance. A left-to-right processing in both recognition and interpretation makes an incremental analysis possible.

I. INTRODUCTION

Though there have been important advances in speech technology throughout the past years only recently some effort has been put in the development of integrated systems for speech understanding. The many highly sophisticated speech recognizers normally demonstrate their capabilities only through figures of recognition accuracy independently from semantic processing. On the other hand side, language understanding has been focused upon in natural language research. Bringing together speech recognition and natural language processing, however, is not easily achieved as the algorithms applied for semantic processing rely heavily on the certainty of the input data. But no current — and also no future — speech recognizer will be able to provide absolutely correct recognition results.

Only some systems applying robust analysis strategies are capable of recognition *and* interpretation of speech. These systems are generally clearly separable into two disjoint processing blocks — statistical word recognition and knowledge based speech interpretation. As a consequence of the different analysis paradigms applied a bottleneck results at the interface between the two components. The most widely used interfacing method is to let the speech recognizer produce some sort of word-hypotheses set for completely given speech data. These hypotheses are then passed to the interpretation component. A true interaction does *not* take place as there is only this unidirectional communication.

Some recent approaches try to integrate recognition and interpretation more closely. But the following aspects that are important to achieve the goal of integrated processing in our opinion are only dealt with partially or are not covered at all:

- Consistent restrictions for valid word sequences have to be used in both recognition and interpretation. This is achieved e.g. by automatically extracting language models from the linguistic knowledge base [11, 14, 1]. In contrast, a linguistic grammar designed by experts can never be guaranteed to be consistent with a statistical language model trained on some corpus.
- If complex linguistic restrictions are applied during the recognition process the information derived should not be lost when producing recognition results. In other words, the result of a complex language model should — and can — also be more complex then the underlying simple sequence of word hypotheses. As far as we know besides our own approach [2] currently no speech recognizer is capable of producing complexly structured hypotheses that can be used directly in speech understanding.
- The control strategy for integrated recognition and interpretation has to be integrated too. So called "tightly coupled systems" have been built mostly as the combination of a natural language parser for LR-grammars and a speech recognizer [8, 6, 4]. Our own approach which yields an incremental analysis strategy is based on the use of complex language models to produce structured recognition results. It will be described in section IV.

This research was supported by the German Ministry of Research and Technology (BMFT) under grant number 01IV102G/7 and by the German Research Foundation (DFG) within SFB 360. Only the authors are responsible for the contents of this publication.

II. LINGUISTIC ANALYSIS

The linguistic analysis is based on a semantic network representation of linguistic knowledge using the ERNEST system [9, 7]. Within a single structured network all the syntactic, semantic, pragmatic and dialogue knowledge needed for the interpretation of speech is stored.

Concepts from this knowledge base that describe coherent linguistic constituents can be transformed into language models automatically as shown in [1]. These form an approximation of the constituent by a regular language. Long term dependencies in discontinuous constituents can *not* be dealt with at this level as their formulation would exceed the capabilities of current HMM-based recognizers.

Besides the capability of representing knowledge for pattern recognition purposes ERNEST also allows its efficient utilization for the analysis. A problem independent control strategy based on the A*-algorithm [7] is supplied. Task dependent extensions of this basic control algorithm constitute the final analysis strategy.

III. ACOUSTIC ANALYSIS

The word recognition module is based on the ISADORA system [12]. It provides highly flexible Markov-model-based speech recognition and the possibility to build structured acoustic descriptions from simple constituents. A rule mechanism is used to represent such complex acoustic models that are capable of describing any regular language of the basic models known to the system. The building of complex HMMs from simpler models by means of these rules does not simply yield a single complex model. The hierarchy of rule constituents is preserved as a hierarchy of HMMs.

The automatic generation of language models mentioned above is therefore able to preserve the structure of linguistic knowledge within the automatically created acoustic models. When applying these in a recognition task the results produced are not simple chains or graphs of word hypotheses. They are also structured according to the associated linguistic model. This makes it easy to map the recognition results on interpretation results that can be used directly by the linguistic analysis process.

IV. INTERACTION OF SPEECH RECOGNITION AND INTERPRETATION

Summarizing the relationships between linguistic and acoustic knowledge we see that the first demand set up in section I for consistency of restrictions is fulfilled. As the structure of knowledge is preserved in the acoustic models produced automatically and in the results obtained from recognition as well the second demand is fulfilled as well. Both properties taken together yield an acoustic knowledge base augmented with abstract constituents that directly correspond to linguistic models as shown in figure 1. The models present in *both* knowledge bases are called *"Semantic Hidden Markov Networks"* (SHMNs) [3] as a semantic network knowledge representation is combined with HMMs.

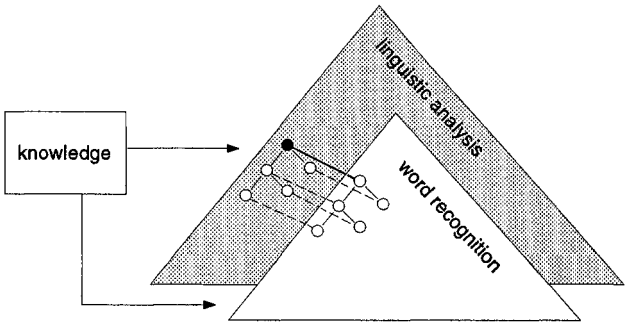

Fig. 1 The integrated knowledge base

However, this modelling technique alone does *not* specify how the combined knowledge base should be applied when recognizing and interpreting speech. As our complex acoustic models are limited to represent regular languages it is not feasible to simply build a model for the whole expected utterance and let the recognizer do all the work. Instead, results of the recognition component should be fed back to the linguistic analysis producing partial interpretations. From these new predictions can be derived dynamically that can be used for constraining the following recognition process.

As the entities modelled in the acoustic knowledge base are not limited to words and as the linguistic processing is not limited to deterministic or probabilistic parsing this interaction scheme substantially extends the integration achieved by coupling of a word recognizer and a LR-parser.

Before the analysis of an utterance it has to be decided which linguistic constituents should be handled by the recognizer. As a dynamic extraction of language models is time consuming and does not provide additional restrictions within our current domain of train-schedule information the corresponding acoustic models are precompiled.

The analysis starts expanding top-down the overall dialogue model until the level of user dialogue steps is reached. With each of these a set of SHMNs is associated able to represent all expected realizations of the corresponding user utterances. Those SHMNs admissable at the beginning of an utterance are passed to the recognizer as *predictions* for frame t_0. No recognition task is started yet. The recognizer collects all predictions starting at the same frame t_0 to a combined task. In the snapshot of the interaction scheme shown in figure 2 these combined tasks are symbolized by the "clouds".

This is necessary as in isolation every SHMN can be found in the speech data though with possibly bad score. The score, however, can only be judged when compared to scores of alternative results. Only the bundling of requests

by the recognizer therefore allows the pruning of extremely bad results thus reducing substantially the search space.

When all predictions starting at the same frame in the input data have been made the linguistic analysis collects *all* the results. When requesting the first acoustic instance the recognition of all predicted models is started in parallel. The complex hypotheses are then passed back to the interpretation component. Due to pruning for some of the predicted SHMNs no results may be computed. The corresponding search path is inadmissable then and discarded from further analysis depicted by the flashes in figure 2.

Within the linguistic search tree the optimal node with respect to a complex scoring function incorporating acoustic and linguistic scores is selected for further expansion according to the underlying A* control algorithm. If the acoustic instances found so far can be combined a new partial interpretation is built. Otherwise, a new set of SHMNs to be predicted is calculated from the total set of predictions for the current dialogue step and the current partial interpretation. If the input data are covered up to frame t the next predictions start at frame $t+1$ and the recognizer takes over again.

This processing scheme is even more flexible then static or dynamic dialogue step dependent language models used e.g. in [10] and [13] respectively. Alternative language models can be active for the same user utterance due to competing predictions on the expected user dialogue step. Though there exists an implicit most general language model for every dialogue step as a collection of all SHMNs allowed this weak model is always restricted to a tighter model by dynamically calculating prediction sets *during* the analysis of an utterance.

As an example let us consider the situation shown in figure 2. The beginning part of the utterance was analyzed successfully covering the word sequence *"I want to go"*. The set of active predictions consists for reasons of simplicity only of *destination*, *time-of-day* and *day-of-week*. The predictions are collected by the recognizer. The combined task is then applied in recognition yielding the hypothesis *"to Heidelberg"* as the best scoring result for *destination*. For *time-of-day* no hypothesis was produced due to pruning and the corresponding search path is discarded. For *destination* and *day-of-week*, however, a partial interpretation is generated from the acoustic instance. Then for the best scoring interpretation *destination* a new set of predictions is calculated consisting of *time-of-day* and *day-of-week* as only a single *destination* is allowed per utterance. The new predictions are passed to the recognizer and processed analogously yielding complex hypotheses that are mapped to analysis results directly.

This processing scheme is continued until the end of an utterance is signalled by the instantation of a special concept. From then on no predictions are allowed any more and the interpretation of the current dialogue step is completed.

A severe problem of *all kinds* of language modelling are portions of an utterance that do not adhere to this model. If no special actions are taken any speech recognition system would map unknown parts of an utterance to the known models of lexicon words or complex constituents regardless of what was really spoken. Therefore, a special model for an unknown constituent or word similar to the ones introduced in [5] is part of *every* prediction set passed to the recognizer. This makes it possible to interpret utterances successfully though possibly only partially that are not conforming to the language model applied. Unknown parts of speech are mapped to a special model and not to some randomly selected garbage instance.

V. RESULTS

We used a speaker independent word recognizer built with the ISADORA-System [12] which was trained on data of 74 speakers that provided 100 domain specific utterances each. The test set consisted of 21 utterances of 4 speakers (3 male, 1 female) who did not provide

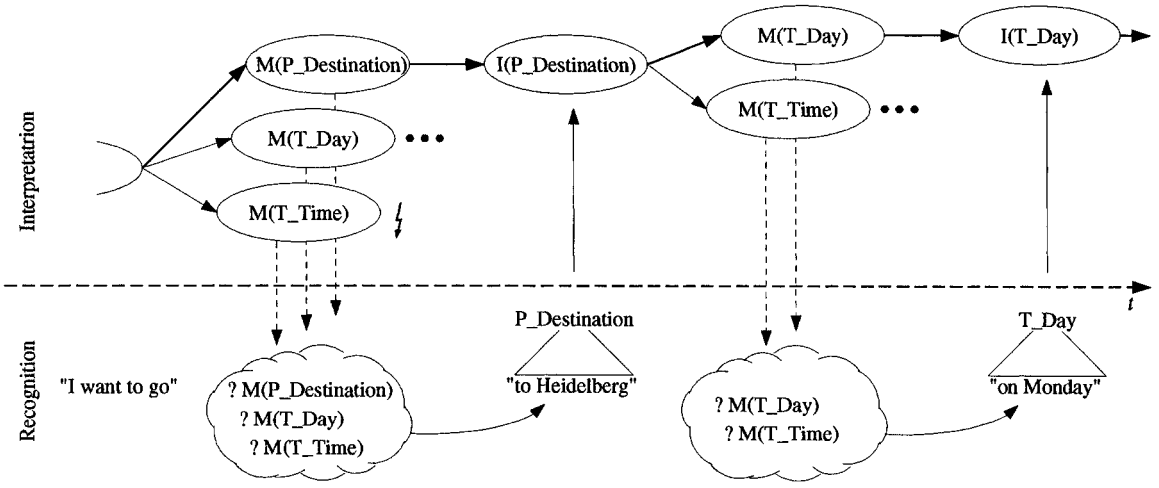

Fig. 2 Snapshot of the interaction between linguistic analysis and word recognition

training material. These test utterances could in principle be analyzed by the language models defined. In this configuration the acoustic model for unknown constituents was merely used as a reference for the acoustic scores of the predicted models. Though our system is capable of interpreting spoken dialogues for evaluation purposes the analysis was applied to a single utterance only.

	Corr	Part	Err	Fail	Tot
Interpretations	79%	1%	13%	7%	100%
Utterances	66	1	11	6	84

Table 1 Evaluation results: interpretations correct (Corr) partial (Part) erroneous (Err) and failed (Fail) and corresponding number of utterances

Table 1 shows that the majority of utterances could be interpreted correctly and *completely*. In these cases all the information relevant for train schedules was computed successfully from the speech data. A partial interpretation may be built if the analysis is terminated prematurely due to a degraded search process. In erroneous results one of the main constituents e.g. *destination* or *time-of-day* were not recognized and interpreted correctly.

The main problems arose from the fact that acoustic scores are not a reliable basis to judge the correctness or even the presence or absence of a specific model. This influenced our first experiments with unknown constituents heavily. Though special modelling and special treatment during the search for an interpretation can be applied it is extremely difficult to find a universal solution neither ignoring most parts of interpretable utterances nor never recognizing an unknown constituent at all.

VI. CONCLUSION

We presented a new technique for building highly integrated speech understanding systems. An integrated acoustic and linguistic knowledge base is established using *"Semantic Hidden Markov Networks"*. Therefore, the restrictions of linguistic knowledge and of the language models used are guaranteed to be consistent. Recognition results can be mapped to partial interpretations directly due to their structural equivalence with the original linguistic knowledge.

Based on the A*–algorithm a control strategy was proposed alternating between prediction and structure building phases processing an utterance incrementally from left to right. Predictions consist of sets of SHMNs and can be derived dynamically during the interpretation of an utterance depending on dialogue context and the current state of the analysis.

Our future work will concentrate on the problem of finding more reliable judgements for incrementally generated recognition results then can be provided by acoustic scores. These should also provide a method for better balancing between hypotheses for known and unknown constituents for improved system robustness.

REFERENCES

[1] G.A. Fink, F. Kummert, and G. Sagerer. Automatic Extraction of Language Models from a Linguistic Knowledge Base. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications*, volume 1, pages 547–550. Elsevier Science Publishers, Amsterdam, 1992.

[2] G.A. Fink, F. Kummert, G. Sagerer, and E.G. Schukat-Talamazzini. Speech Recognition Using Semantic Hidden Markov Networks. In *Proc. European Conf. on Speech Technology*, pages 1571–1574, 1993.

[3] G.A. Fink, F. Kummert, G. Sagerer, E.G. Schukat-Talamazzini, and H. Niemann. Semantic Hidden Markov Networks. In *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 919–922, Banff, Canada, 1992.

[4] David Goddeau. Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems. In *Proc. Int. Conf. on Spoken Language Processing*, pages 321–324, Banff, Canada, 1992.

[5] A. Jusek, H. Rautenstrauch, G. A. Fink, F. Kummert, G. Sagerer, J. Carson-Berndsen, and D. Gibbon. Detektion unbekannter Wörter mit Hilfe phonotaktischer Modelle. In *Mustererkennung 94, 16. DAGM-Symposium Wien, Informatik aktuell*. Springer-Verlag, Berlin, 1994. to appear.

[6] Kenji Kita and Wayne H. Ward. Incorporating LR Parsing into SPHINX. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 269–272, 1991.

[7] F. Kummert, H. Niemann, R. Prechtel, and G. Sagerer. Control and Explanation in a Signal Understanding Environment. *Signal Processing, special issue on 'Intelligent Systems for Signal and Image Understanding'*, 32:111–145, 1993.

[8] Hy Murveit and Robert Moore. Integrating Natural Language Constraints into HMM-based Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 573–576, 1990.

[9] H. Niemann, G. Sagerer, S. Schröder, and F. Kummert. ERNEST: A Semantic Network System for Pattern Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:883–905, 1990.

[10] 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.

[11] Fernando Peireira. Finite-State Approximations of Grammars. In *Speech and Natural Language Workshop*, pages 20–25, Hidden Valley, Pennsylvania, 1990. Morgan Kaufmann.

[12] E.G. Schukat-Talamazzini and H. Niemann. Das ISADORA-System — ein akustisch-phonetisches Netzwerk zur automatischen Spracherkennung. In *Proc. 13. DAGM-Symposium*, pages 251–258. Springer, Berlin, 1991.

[13] S.R. Young, A.G. Hauptmann, W.H. Ward, E.T. Smith, and P. Werner. High Level Knowledge Sources in Usable Speech Recognition Systems. *Communications of the ACM*, 32(2):183–193, 1989.

[14] Victor Zue, James Glass, David Goodine, Hong Leung, Michael Phillips, Joseph Polifroni, and Stephanie Seneff. Integration of Speech Recognition and Natural Language Processing in the MIT VOYAGER System. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 713–716, 1991.