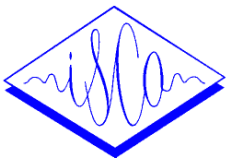

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

SEMANTIC HIDDEN MARKOV NETWORKS*

G. A. Fink, F. Kummert, G. Sagerer, E. G. Schukat-Talamazzini‡, H. Niemann‡

Universität Bielefeld, Technische Fakultät, AG Angewandte Informatik
Postfach 100 131, D-4800 Bielefeld 1, Federal Republic of Germany
Tel.: +49–(0)521–106–5329, Fax: +49–(0)521–106–2992

‡Universität Erlangen-Nürnberg, IMMD 5 (Mustererkennung)

ABSTRACT

Although much effort has been put into speech understanding systems there still exists a rather wide gap between acoustic recognition and linguistic interpretation. We propose a formalism for an extremely close interaction of acoustic recognition and higher level analysis. Instead of a strict *horizontal* interface at the level of hypothesized word sequences or lattices, a *vertical* interface to the acoustic component is used that can be accessed from linguistic concepts of any degree of abstraction. As the linguistic knowledge is represented in the formalism of *Semantic Networks* and acoustic recognition is based on *Hidden Markov Models* the close interaction between the two components was termed *Semantic Hidden Markov Networks*.

10.21437/ICSLP.1992-282

1 INTRODUCTION

Because of the high degree of uncertainty in the recognition of spoken language it is very important to exploit any possible predictions and restrictions to guide acoustic analysis as well as linguistic interpretation. Within traditional systems the possibilities of interaction between linguistics and acoustics are very limited because of the use of completely different formalisms that lack a combination. *Semantic Hidden Markov Networks (SHMNs)*, however, serve to combine closely the low and high levels of analysis within a speech understanding system. Using this new joint formalism it is possible to derive acoustic constraints from linguistic restrictions as well as to instantiate abstract linguistic concepts by simply detecting a SHMN in the speech signal. This is based on the fact that a mapping of complex linguistic structures onto acoustic models and vice versa is made. A SHMN is just an acoustic representation of a linguistic constituent that preserves the internal hierarchical structure and reflects the restrictions on the basic components as tightly as possible, i.e. the combination of a language model, an acoustic model, and the reference to a linguistic concept. Traditionally the most abstract linguistic element that could be represented by an acoustic model was a single word. SHMNs extend this basic technique to linguistic concepts of any degree of abstraction. As SHMNs can be created dynamically, predictions made during linguistic analysis can be incorporated directly into the acoustic recognition process. This method of feedback is even more flexible and general than the use of various language models for e.g. different dialogue steps.

2 LINGUISTIC ANALYSIS

The formalism of a semantic network is used to represent linguistic knowledge as specified by the ERNEST language [4]. Important features of ERNEST are the existence of *modified concepts* that represent concept descriptions incorporating constraints arising from analysis and *adjacency matrices*. The latter serve to describe the possible well formed sequences of a concept's components.

Syntactic, semantic, and pragmatic knowledge is stored within a single semantic network using the uniform representation language. This description of linguistic knowledge is however clearly separable into distinct levels of abstraction.

- The *hypothesis* level forms the traditional interface between acoustic recognition and linguistic analysis.
- Concepts describing the structure of syntactic constituents form the *syntax* level. However, a sentence grammar is not used.
- Deep-case theory [1] provides the framework used to describe the meaning of syntactic components on the *semantics* level using problem independent noun and verb frames.
- Task specific knowledge is stored in the concepts of the *pragmatics* level. Semantic descriptions are restricted to their specific use in the task domain of train information. Additionally, concepts interfacing with a database are provided.
- On the *dialogue* level the possible user and system utterances and the relations among them are described. Actually the modelling allows a first request of the user followed by system requests for detail or confirmation until a database query can be made and a train schedule is obtained.

To use extensively the expectations of the linguistic knowledge base a partial interpretation is not extended by a sequential processing of the speech signal but on the basis of structural relations. New hypotheses are accepted for a partial interpretation if they satisfy the restrictions resulting from the constraint propagation process within the semantic network. The aim of the linguistic analysis is the instantiation of a concept representing a user dialog step. Due to the uncertainty of the word recognition module and due to the wide variations of utterances in spoken language neither a strictly data–driven nor a pure model–driven strategy seems to be promising when working with an acoustic module that uses no task dependant information for word recognition. Therefore, currently we use a strategy working both on the acoustic data as well as on the expectations derived from the linguistic model [3]. For a goal directed search of the best fitting interpretation the A*-algorithm with a judgment vector is used reflecting the compatibility, the quality, the reliability and the relevance of a partial interpretation [4].

3 ACOUSTIC ANALYSIS

The acoustic part of the analysis is covered by the ISADORA system [5] which provides highly flexible speech recognition based on Hidden Markov Models (HMM).

* This research was partially supported by the German Ministery of Research and Technology (BMFT) under grant number 01IV102A0. Only the authors are responsible for the contents of this publication.

ISADORA consists of modules for cepstral feature extraction, for optional hard or soft gaussian vector quantization, and for beam-search driven training and Viterbi decoding (see Figure 1). Presently, the HMM emission probabilities can be chosen to be either of discrete, (single gaussian) continuous, or semi-continuous type.

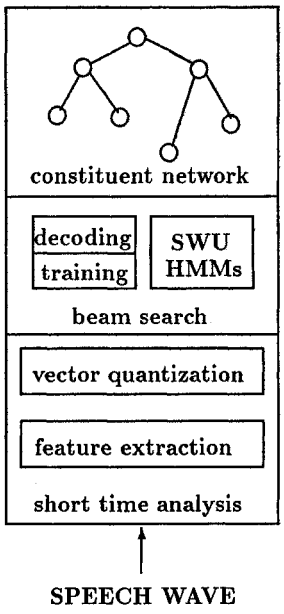

Fig. 1 The overall architecture of ISADORA

Structural knowledge is represented in ISADORA by means of a huge constituent network the nodes of which correspond to different speech concepts, e.g. phonetic units, morphemes, words, sentences, vocabularies, finite-state grammars, and others. The network is restricted to form an acyclic graph, and each node is characterized by its name, its type, and a list of successor nodes. Each node is acoustically represented by an HMM related to the successors' models as specified by the node type:

- **A-nodes** are the *atoms* of the network. They are provided with a dedicated HMM in order to acoustically represent the corresponding speech unit.
- **S-nodes** denote the *sequential* concatenation of the successor's models to form a larger HMM.
- **P-nodes** serve to represent *paradigms*, i.e. the choice between alternative speech concepts. The realization is by connecting in *parallel* the successor's HMMs.
- **R-nodes** denote an arbitrary repetition a single successor concept. This type is implemented by looping the successor model.
- **F-nodes** provide the most general means of hierarchical model construction. The F-node model is a *finite-state network* obtained by interconnecting the successor's models according to a particular adjacency matrix.

Note that within the ISADORA network formalism not merely finite-state structures as well as regular expressions can be defined. All of these constructions may be combined or nested in any order and depth. For a concrete speech understanding application, the above techniques are exploited in several ways:

- on the phonetic level, even sophisticated subword speech unit modelling approaches, (generalized triphones or context-freezing units, see [6]) can be implemented easily,
- on the morphological level, for instance, compound words and numerals can be represented in a very economical way,
- grammars can be (pre-)defined to restrict the acoustic search to well-formed sentence structures, or even to introduce situation-dependent top-down constraints,
- and finally, the analysis task to solve can be specified by simply pointing to an appropriately structured ISADORA concept.

  As an example, connected word recognition is obviously modelled by an R-node repeating a P-node which in turn bundles a certain set of word nodes that form the vocabulary.

  It is very important to note the fact that such complex analysis task concepts can be created "on the fly", i.e. *during* the understanding process.

At the decoding stage, ISADORA performs a beam-search driven Viterbi algorithm on the HMM that corresponds to the analysis node trying to find a best-fitting time alignment with the acoustic input. The result of this procedure is a nested instance structure, being essentially isomorphic to the analysis concept, but additionally providing the scores and time boundaries resulting from the time alignment.

4 ACOUSTIC-LINGUISTIC INTERFACE

4.1 The Traditional Approach

The most widely used technique of interfacing between acoustic recognition and linguistic interpretation is to pass hypothesized words from the acoustic to the linguistic component (Fig. 2). These word hypotheses are mostly embedded in word lattices or a certain number of the best fitting word sequences. Though often some a priori knowledge is used in form of statistical or finite state language models there is never an interaction in the opposite direction, i.e. the linguistic component is allowed to *verify* words or sequences but not to completely *hypothesize* them. Furthermore the knowledge base of the linguistic and acoustic analysis are *completely distinct.* The acoustic knowledge base — if any — does never really reflect linguistic knowledge, because all approved techniques for the generation of language models use statistical approaches as for example bi- or trigramms. The parameters of theses statistic models, however, can not be calculated reliably in many cases, because of the lack of significantly large text corpora from which the probabilities are estimated.

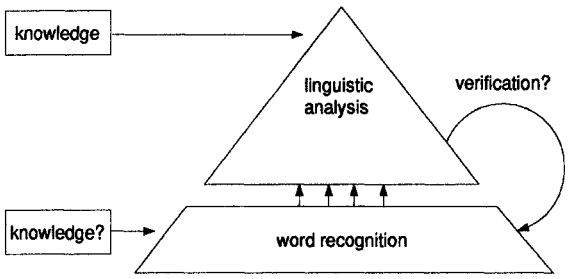

Fig. 2 Traditional interfacing method between the acoustic and linguistic component

4.2 Static SHMNs

The main idea of SHMNs is to perform a mapping of linguistic structures onto acoustic models. This mapping preserves the constituent structure of the linguistic model because also HMMs can form structured acoustic descriptions within ISADORA. The restrictions on sequential relations between components and linguistic categories matching them are reflected as tightly as possible. As the semantic network representation of linguistic knowledge is more powerful

than the regular capabilities of SHMNs the later will in general cover a superset of the original description. Though, this kind of generalization weakens the correspondence between the linguistic concept and the associated SHMN it may also be desired to drop some of the linguistic restrictions for to obtain more efficient acoustic models.

Automatic Generation of Language Models: As it is desirable to use the same knowledge sources for both acoustic constraints and linguistic interpretation we developed a method for automatically extracting language models from the linguistic knowledge base. This process uses the control structures provided with the semantic network language ERNEST to expand a linguistic model into an acoustic network (for a detailed description see [2]).

In principle this method acts very similar to a model driven linguistic analysis procedure. A given linguistic prediction about some constituent stored in a modified concept is recursively expanded into its parts and concretes while restrictions are propagated and the sequential relationships between constituents that are stored in adjacency matrices are used to form the acoustic network. It is important to note that the structure of the resulting SHMN is not "flat" but corresponds to the concept hierarchy describing the expanded model.

For the task domain of train information a very illustrative example is the expansion of the concept P_DESTINATION into a SHMN that restricts the word sequences accepted for the instantiation of this concept to the utterance scheme "to <some town>". Figure 3 shows on the left hand side the simplified concept structure of P_DESTINATION and on the right the ISADORA-rules that are created. The restriction of the preposition and reduction of the noun phrase to the mere nucleus result from the restrictions that are imposed by the pragmatic concept.

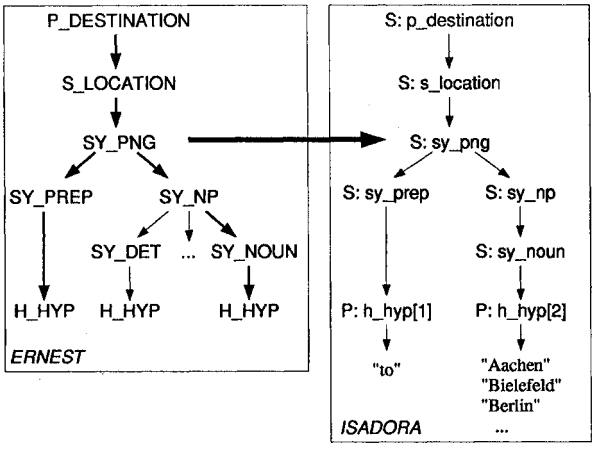

Fig. 3 Mapping of a pragmatic concept onto an acoustic model

For to obtain the most efficient SHMNs those concepts have to be chosen that can impose the tightest restrictions on their realizations in the task domain while still being tractable by acyclic constituent structures. These restrictions should include sequential relationships. From the structure of the linguistic knowledge base outlined in section 2 it can be seen that the concepts of the pragmatics level meet those needs best because they reflect syntactic structures restricted to a specific use within the task domain. A general noun phrase would, however, expand into a huge SHMN with only very weak predictive power because domain dependant restrictions can not be exploited.

Use of Highly Abstract Acoustic Concepts in Linguistic Analysis: When concentrating on the first utterance of an informational dialogue about train connections the expected pragmatic constituents are the best candidates for automatically created language models. An arbitrary sequence of P_TRAVELLER, P_DESTINATION, P_DEP_PLACE, P_TIME and a concept covering directional and modal verbs would already describe the most common request schemes, e.g. *"I want to go from Bielefeld to Frankfurt this evening."*. A collection of the SHMNs corresponding to the above mentioned concepts can in a first step be used as a traditional language model of the utterance. The difference between such an off-line language model and a SHMN is the fact that the later can be accessed directly from the linguistic component while the first serves just to constrain the possibly generated word hypotheses. When trying to instantiate a certain linguistic concept it is therefore no longer necessary to perform some kind of model expansion until a communication with acoustics via word hypotheses is possible. Once given the relation between linguistic and acoustic structures the analysis process can just calculate a score for the SHMN of the specified constituent obtaining a structured instance of the acoustic model. The corresponding linguistic interpretation, however, is not yet available. For intermediate interpretations of the speech signal it does not have to be calculated. Only when the sequence of SHMNs matching the speech data best is found a remapping of acoustic models onto linguistic constituent structures has to be made. Because of the structural equivalence between SHMNs and linguistic concepts this can be achieved with only minor effort.

If due to the differences in expressive power between SHMNs and the original linguistic descriptions an inconsistency is detected during the remapping process the interpretation found has to be rejected. The A*-driven analysis is then to find an alternate interpretation.

Figure 4 shows the architecture of a speech understanding system using the method of communicating between acoustics and linguistics outlined above. We termed this version *static* SHMNs because the abstract acoustic models are only *accessed* at analysis time and *not created.* This has to be done in a prephase of the analysis. Note that the knowledge base used for both generation of SHMNs and linguistic analysis is *identical.*

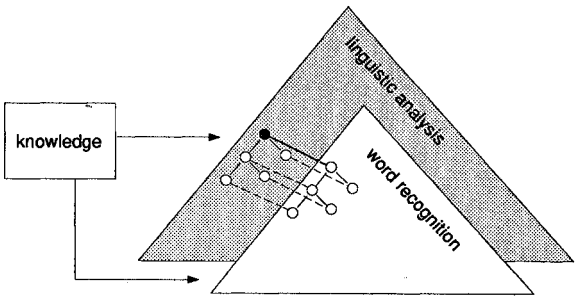

Fig. 4 Acoustic-linguistic interface using static SHMNs

4.3 Dynamic SHMNs

The use of static SHMNs makes a close interaction between linguistics and acoustics possible. Restrictions, however, that emerge during the understanding process can not be used to constrain acoustic recognition. An extension of the SHMN technique to *dynamic* models integrates the creation of language models into the analysis. When trying to instantiate some linguistic concept *not* a precompiled SHMN is used but instead the required model is created *dynamically* and can then be detected by the acoustic component.

Figure 5 shows the extended system architecture incorporating *dynamic SHMNs.* The static component is still present to serve as a basic collection of acoustic models. It is however extended by newly created SHMNs that are based on existing models and incorporate

the most specific predictions that were calculated from the results of the analysis done so far. Dynamic SHMNs are the result of a *feedback* of predictive information from linguistics to acoustics.

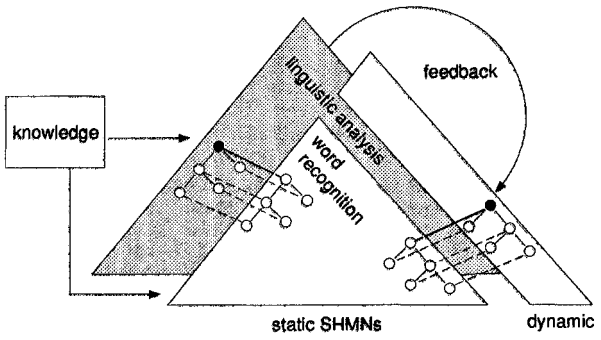

Fig. 5 Acoustic-linguistic interface using dynamic SHMNs with feedback of intermediate results to newly created acoustic models

5 FIRST RESULTS

The first evaluations using SHMNs in the understanding of spoken language concentrated on the problem of how to initiate the instantiation on a SHMN from the linguistic analysis.

Constituent Spotting: Our existing speech understanding system provides structure oriented analysis that alternates between data and model driven phases. Therefore, in some experiments we tried to let acoustics cope with the data and to just predict SHMNs for linguistic constituents by the analysis. However, as those predictions can not incorporate sequential relations between different SHMNs if computed separately the results of "spotting" the SHMNs in the speech data showed the same uncertain results as the spotting of word hypotheses. Furthermore, from the mere comparison of acoustic scores it can not be decided reliably whether the instantiation of a SHMN was successful or just forced.

Dialogue Step Dependant SHMNs: To circumvent the problem arising from the lack of contextual information when positioning SHMNs we decided to derive SHMNs describing complete utterances from the knowledge base. As at present there exist linguistic constituents that can not yet be transformed into efficient acoustic models automatically (the resulting HMM structures are still about 100 times larger than "normal" models), the dialogue step SHMNs were reduced in complexity by hand. We obtained five different models: REQUEST, DECISION, P_DESTINATION, P_DEP_PLACE and S_TIME. REQUEST and DECISION consist of an arbitrary sequence of the SHMNs corresponding to P_TRAVELLER, P_DESTINATION, P_ROUTE, P_DEP_PLACE, S_TIME and SY_VERB preceeded by an optional greeting phrase or a required "yes" or "no" respectively. We used an acoustic recognizer based on HMMs with continuous emission probabilities that was trained using 500 sentences from the domain of train information uttered by four male speakers. Using 20 dialogues of a speaker that belonged to the training set we compared recognition results to an approach which used no language model at all and a lexicon of 1081 words. Without the use of SHMNs a word accuracy of 75% and a sentence recognition rate of 51% were achieved with *the same* acoustic recognizer but a different analysis task. Using the appropriate dialogue step dependant SHMN for the analysis of the utterances a word accuracy of 94.5% was reached with 81% of the utterances being recognized completely correct. As the error rate was cut by nearly 80% it can be expected that even more complex SHMNs covering more complex dialogue utterances still allow for a significant improvement of acoustic recognition.

It should be mentioned that the instantiation of a dialogue SHMN yields not only a best fitting word sequence but also a sequence of constituent hypotheses. The reinterpretation of these hypotheses we are currently working on is the missing feature for the realizations of a speech understanding system based on static SHMNs.

6 CONCLUSION

We presented a new technique of interfacing between acoustic and linguistic analysis. It offers two major advantages over the traditional passing along of word hypotheses:

- Concepts of linguistic knowledge can be modeled using almost equivalent structures of the acoustic component, i.e. a mapping of structural knowledge from linguistics to acoustics and vice versa is possible.
- Structural interaction is *not* limited to predefined concepts.

From the discussion of our first evaluation experiments it can be seen, that SHMNs will probably yield good results when incorporated into a left-to-right linguistic analysis system. There the left acoustic context for the SHMN to be instantiated is always given and restricts the possible matching models and the predictive power of SHMNs can be exploited substantially. Although the results presented are only preliminary it can be stated that the interaction of acoustics and linguistics via SHMNs improves the recognition and interpretation of spoken language in the following ways:

- Acoustic and linguistic knowledge are consistent.
- Interfacing of high and low level processing can be done by passing entire constituent hypothesis from acoustics to linguistics.
- Dynamic SHMNs allow for acoustic predictions derived from the current state of linguistic analysis.

Our future work will concentrate on the problem of reinterpreting acoustic structures that were found to match the speech signal best. Furthermore, the dynamic aspect of SHMNs shall be incorporated into a left-to-right linguistic analysis system.

References

[1] Ch. Fillmore. A case for case. In E. Bach and R. T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York, 1968.

[2] G. A. Fink, F. Kummert, and G. Sagerer. Automatic extraction of language models from a linguistic knowledge base. In *6th European Signal Processing Conference*, Brussels, Belgium, 1992.

[3] F. Kummert. *Flexible Steuerung eines sprachverstehenden Systems mit homogener Wissensbasis*. PhD thesis, Technische Fakultät der Universität Erlangen-Nürnberg, 1991.

[4] G. Sagerer and F. Kummert. Knowledge based systems for speech understanding. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, pages 421–458. NATO ASI Series F, Vol. 46, Springer-Verlag, Berlin, 1988.

[5] E. G. Schukat-Talamazzini and H. Niemann. Das ISADORA-System – ein akustisch-phonetisches Netzwerk zur automatischen Spracherkennung. In B. Radig, editor, *Mustererkennung 1991*, volume 290 of *Informatik Fachberichte*, pages 251–258, Berlin, Heidelberg, New York, 1991. Springer Verlag.

[6] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Acoustic modelling of subword units in the isadora speech recognizer. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 577–580, San Francisco, 1992.