A ONE-PASS SEARCH ALGORITHM FOR UNDERSTANDING NATURAL SPOKEN TIME UTTERANCES BY STOCHASTIC MODELS

Josef G. Bauer[1], Holger Stahl[2], Johannes Müller[2]

[1] Siemens AG
Corporate Research and Development
Dept. ZFE T SN 53
D-81730 Munich, Germany
email: Josef.Bauer@zfe.siemens.de

[2] Institute for Human-Machine-Communication
Munich University of Technology
Arcisstrasse 21,
D-80290 Munich, Germany
email: {sta,mue}@mmk.e-technik.tu-muenchen.de

ABSTRACT

A system for understanding time utterances spoken in German language is presented. Stochastic models contain the knowledge in the semantic, syntactic and acoustic-phonetic levels. An adequate semantic representation allows the integration of these models within a one-pass Viterbi search. The simultaneous use of all knowledge sources for the search procedure results in the smallest possible search space for the determination of the most probable semantic content accurately following the Bayes classification rule. Both the recognition accuracy and the computing speed facilitate a realistic application.

Keywords: speech recognition, language understanding, spoken man-machine-dialogue, stochastic models, one-pass search, representation of syntactic and semantic knowledge

1. INTRODUCTION

Using speech as a communication medium in technical systems, the syntax and the semantics of the spoken utterances are often strongly constrained within a certain domain. The research system described in this paper exclusively handles utterances that describe a particular time of day, spoken in German language. Their semantic content is defined as a cardinal number S counting the number of minutes elapsed since midnight, with $0 \le S \le 1439$. S can also be expressed by a tupel (h, m) of two numbers with $0 \le h \le 23$ and $0 \le m \le 59$:

$$S = S(h, m) = 60 \cdot h + m \qquad (1)$$

In this case, h absolutely marks the full hour and m the amount of minutes, which have passed additionally.

An example for the word chain of an utterance within this domain is "*zwei nach dreiviertel acht*" ("*two past a quarter to eight*"). According to the definition above, the semantic content of this word chain is $S(h = 7, m = 47) = 467$. The same semantic content $S = 467$ can also be expressed by other word chains like "*sieben uhr siebenundvierzig*" ("*seven fourtyseven*") or "*dreizehn minuten vor acht*" ("*thirteen minutes to eight*").

The desired capability of the system is the extraction of the semantic content S from such 'time utterances' available as continuous speech from an unknown speaker.

2. MAXIMUM-A-POSTERIORI DECODING

Speech understanding can be interpreted as mapping a sequence of observation vectors O (i.e. the preprocessed speech signal of the utterance) to its semantic content S. The maximum-a-posteriori decoding criterion is used for finding the most likely semantic content

$$S_E = \underset{S}{\operatorname{argmax}}\, P(S|O) . \qquad (2)$$

Applying the Bayes' inversion formula and neglecting $P(O)$, which is constant within the maximization, this equation can be written as

$$S_E = \underset{S}{\operatorname{argmax}}\, [P(O|S) \cdot P(S)] . \qquad (3)$$

The a-priori probability $P(S)$ of a certain semantic content can be estimated directly from a limited training corpus, whereas this is not possible for the conditional probability $P(O|S)$ due to the high variety of S and O [3]. By the use of the additional representation levels *close-to-word-level semantic representation U* and *word chain W*, it is possible to split the modelling problem into individual parts, which can be solved separately and expressed by first order conditional probabilities. Assuming statistical independence of these probabilities, $P(O|S)$ can be written as

$$P(O|S) = \sum_U \sum_W \left[P(O|W)\, P(W|U) P(U|S)\, P(S)\right] . \qquad (4)$$

Considering only those values of W and U, which maximize (decoding by the Viterbi algorithm) the joint probability

$$P(O, W, U, S) = P(O|W) \cdot P(W|U) \cdot P(U|S) \cdot P(S) , \qquad (5)$$

we obtain the classification rule

$$S_E = \underset{S}{\operatorname{argmax}}\, \max_U \max_W P(O, W, U, S) . \qquad (6)$$

The top-down decoding (fig. 1) of the semantic content S_E can be described as a synthesis of hypotheses at the different representation levels [5] followed by a detection of the hypotheses with the maximum joint probability according to eq. (6). The hypotheses originated by the 1st order semantic model, the semantic generator and the word chain generator occur with the corresponding joint probabilities $P(S)$, $P(U, S)$ and $P(W, U, S)$. Directly implementing the top-down decoder following fig. 1, the probability $P(O, W, U, S)$ would have to be computed for each word chain W with the corresponding $P(W, U, S) \neq 0$ explicitly. Chapters 4 and 5 describe the methods employed to avoid these explicit computations for the implemented system.

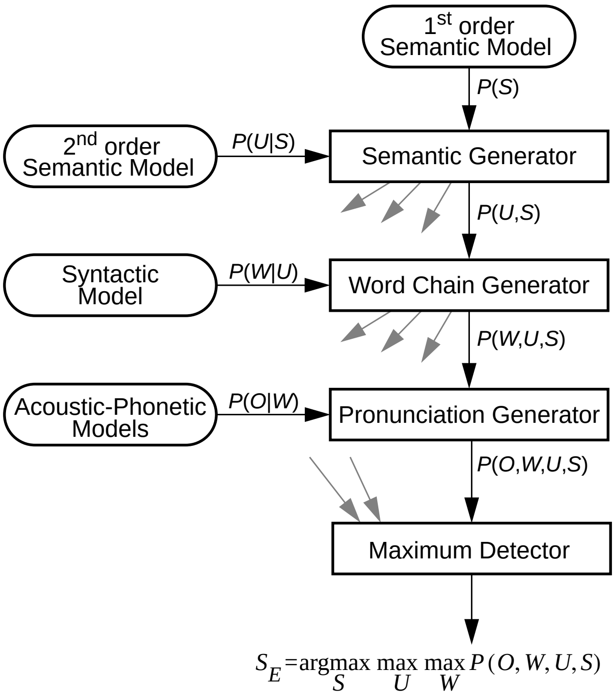

Figure 1: Search for the most probable semantic content by generating hypotheses in a 'top-down' strategy

3. STOCHASTIC MODELS

The *1st order semantic model* specifies the a-priori probability $P(S)$ for the occurrence of the semantic content S within a particular domain of interest. $P(S)$ can either be estimated from a training corpus or it can be specified using semantic expert knowledge, available in a certain human-machine-dialogue situation. Both for easily designing the model and for efficiently decoding the utterance, it is advantageous to use a discrete probability distribution as shown exemplary in Fig. 2:

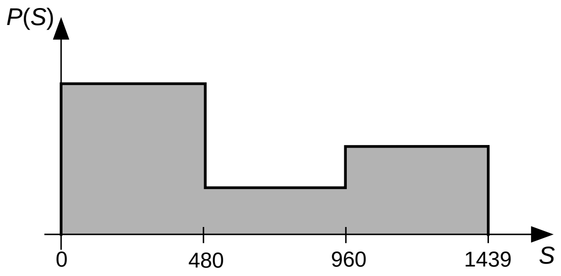

Figure 2: Discrete 1st order semantic model

The *2nd order semantic model* contains the conditional probabilities $P(U|S)$ for the occurrence of the semantic representation U given S. The close-to-word-level semantic representation U of a time utterance is defined as a triple (x, y, z) with x denoting the hours, y denoting the quarters of an hour and z denoting the minutes. The value x $(0 \leq x \leq 24)$ indicates an absolute time, y $(-3 \leq y \leq 1)$ and z $(-29 \leq z \leq 59)$ are data relative to x. For the above example of the German word chain W: "*zwei nach dreiviertel acht*", the close-to-word-level semantic representation $U = (8, -1, 2)$ is obtained.

There exists an unambiguous mapping function to determine the semantic content S from a given U:

$$S(U) = S(x, y, z) = (60 \cdot x + 15 \cdot y + z) \bmod 1440 \qquad (7)$$

The conditional probability $P(U|S)$ can be estimated in a parametric form from training data using histogram methods and parameter smoothing. Due to the lack of an adequate amount of training material we applied this technique on the test corpus to estimate $P(U|S)$ as well as $P(S)$. It is assumed, that $P(U|S) = 0$ for $S(U) \neq S$ and that $P(U|S)$ is independent of the value x. Fig. 3 shows an example for a possible discrete probability distribution, only depending on the values y and z.

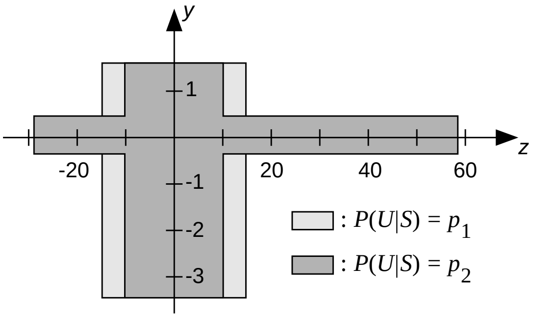

Figure 3: Discrete 2nd order semantic model

The *syntactic model* specifies the conditional probabilities $P(W|U)$ for the occurrence of the word chain W given U. The syntax of pure time utterances can be expressed by a regular grammar G which is equivalent to a finite state automaton A, so that the formal languages $L(G)$ and $L(A)$ are equal [4] [7]:

$$L(G) = L(A) \qquad (8)$$

The automaton can be seen as a transformer $U(W)$, which accepts a sequence of input symbols and produces a sequence of output symbols. In our case, the input is the word chain W, the output causes allocations to the components x, y and z of the close-to-word-level semantic representation U. Since the examined domain is still comprehensible, the syntactic model is originated by 'expert' knowledge instead of training it. Thus, it is assumed that $P(W|U)$ represents a uniform distribution for all word chains which are valid for a given U:

$$P(W|U) = \begin{cases} 1, & W \in L(A) \wedge U(W) = U \\ 0, & \text{else} \end{cases} \qquad (9)$$

Fig. 4 shows a syntactic model automaton consisting of the start/end state, 8 inner states and 11 edges, which is able to process a limited set of German time utterances. The model implemented in our system to understand arbitrary German time utterances is about five times larger.

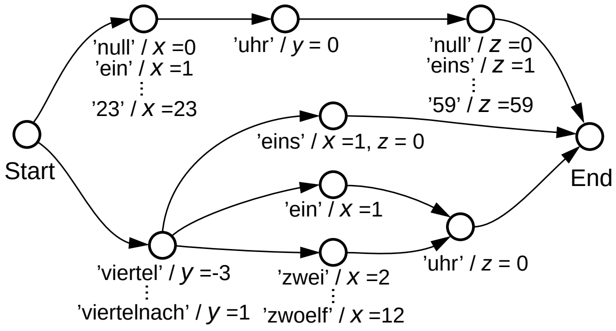

Figure 4: Simplified syntactic model, consisting of 10 states, 11 edges and possible input/output symbols

Every pair of input/output symbol positioned under the states of the graph in fig. 4 represents one possible transition for each edge ending in the specific state. Each transition accepts one input symbol, i.e. one word out of *W*, written left of the slash '/'. The transition optionally allocates a value to one or more components *x*, *y* and *z* of *U*, written right of the slash.

It must be guaranteed, that:

- Every valid word chain is accepted by the automaton, i.e. the automaton is passed from 'start' to 'end'.
- Every path from 'start' to 'end' causes exactly one unambiguous allocation to each of the components of *U*.

The *acoustic-phonetic models* serve for calculating the conditional probability $P(O|W)$ of emitting an observation sequence *O* given the word chain *W*. We use phoneme-based, continuous Hidden-Markov-Models, which were trained from a multi-speaker phonetically balanced speech data base. Several of these phoneme models are then assembled to entire word models by using a pronouncing dictionary. These acoustic modelling technique has proved to be powerful in many of the present speech recognition systems [8] [2] and was adopted unchanged.

4. INTEGRATION OF THE SYNTACTIC AND THE SEMANTIC MODELS

For the search procedure, a language model is required, which contains all information from the 1st and 2nd order semantic models and from the syntactic model. For keeping the computing and memory effort manageable during the search, hypotheses with the same probability $P(U, S)$ can be regarded in common under certain conditions. The arising *hypothesis-clusters* are described in a parametric form by sets Ψ_i of values *U* where X_i, Y_i and Z_i are sets of values *x*, *y* and *z:*

$$\begin{aligned} \Psi_i &= X_i \times Y_i \times Z_i \\ &= \{ (x, y, z) = U | (x \in X_i \wedge y \in Y_i \wedge z \in Z_i) \} \end{aligned} \qquad (10)$$

In our implemented system, one single parametric description Ψ_i represents up to 2225 explicit hypotheses for *U*. With this fundamental restriction, it is possible to derive an automaton A_i from a syntactic model described in chapter 3 that the following relation holds:

$$W \in L(A_i) \Leftrightarrow U(W) \in \Psi_i \qquad (11)$$

The automaton A_i must only contain those transitions, which generate output symbols allocating values *x*, *y* and *z* that follow the restriction $x \in X_i \wedge y \in Y_i \wedge z \in Z_i$.

The *language model* is composed by paralleling all these reduced syntactic models, which have the entry probability $P(U, S) = P(\Psi_i)$, respectively. Fig. 5 shows an example for a language model build from three automata, each derived from the automaton in fig. 4, required for three hypothesis-clusters.

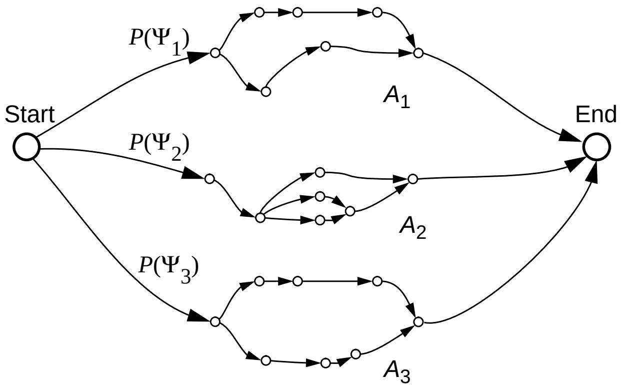

Figure 5: Example language model, consisting of three paralleled automatons A_1, A_2 and A_3

5. SEARCH PROCEDURE

To avoid generating all hypotheses on the representation levels in fig. 1 explicitly, the most likely semantic content S_E is computed using the Viterbi search algorithm. Two different approaches were investigated:

5.1 Two-pass search

At the first stage, the *N* most likely hypotheses for the word chain are determined. The result of this stage is either a word lattice or just the most likely word chain. We only consider the case $N = 1$, i.e. the most probable word chain W_E is found using a maximum likelihood classification:

$$W_E = \underset{W}{\operatorname{argmax}}\, P(O|W) , \qquad (12)$$

The language model is taken into account exclusively at the second stage:

$$S_E = \underset{S}{\operatorname{argmax}}\, \underset{U}{\max} \left[P(W_E|U) \cdot P(U|S) \cdot P(S) \right] \qquad (13)$$

This approach can not constitute an optimal solution for finding the most probable semantic content *S*. It is not guaranteed, that the word chain maximizing $P(O, W, U, S)$ according to eq. (6) is among the hypotheses passed over to the second stage, since *N* is limited.

5.2 One-pass search

The language model as well as the acoustic-phonetic models are taken into account in a one-pass search procedure [6].

The actual search space results from combining the word HMMs according to all transitions in the language model. With a medium number of 30 states per word HMM the theoretical maximum search space reaches a size that requires a dynamic organisation of the states. Using a beam search technique [8], the number of grid points that have to be processed can be kept in a range comparable to the one resulting from a bi- or unigram language model. The simultaneous use of all knowledge sources during the one-pass search procedure results in a high accuracy what hypotheses can be pruned at the respective moment. In our implementation, an adequate pruning strategy leads to a reduction of computing time by the factor two without any effect to the recognition accuracy.

Making use of the output symbols produced by the language model, the most likely semantic content can be determined without any backtracking procedure. By keeping track of all allocations to x, y and z along all paths, S_E is computed applying eq. (7) to the values along the most likely path. Neglecting the effects of the pruning, the described one-pass search decodes the most probable semantic content according to eq. (6).

6. EXPERIMENTAL RESULTS

The performance of the system was evaluated using 1023 utterances spoken by 20 different speakers, recorded in a low noise office environment. The speakers were instructed to talk in their natural manner with normal pronunciation.

	$P(U,S) = const$	$P(S)$ and $P(U\|S)$ trained
two-pass search	28 %	28 %
one-pass search	78 %	87 %

Table 1: Percentage of correctly assigned semantic contents under different conditions

Chosen $P(U,S) = const$ for all hypotheses U, 28% of all semantic contents were correctly assigned using the two-pass search. For the one-pass search the share of correctly assigned sentences was 78%. The considerable lower recognition rate for the two-pass search is caused by the lack of semantic and syntactic knowledge, when determining the spoken word chain. This leads to wrong or even void word chains, passed over to the second stage.

Semantic models for $P(S)$ and $P(U|S)$ adapted to the semantic contents in the test corpus increased the recognition rate up to 87% in the case of the one-pass search. It is obviously clear why this improvement of the language model had no effect when applying it after determining only the most likely word chain according to eq. (12) in the two-pass search.

Certain situations in spoken dialogue systems, e.g. a voice controlled booking system, lead to an extreme low number of possible hypotheses for S (times of departures, etc.). Using semantic knowledge of this kind, it is even possible to improve the accuracy of the system up to 100 % [1].

7. CONCLUSIONS

The use of stochastic models for representing semantic knowledge in connection with an efficient syntactic modelling has been proven to be an important step towards a realistic applicability of speech understanding systems. The described system can be employed not only for time utterances, but in principle for all utterances whose language can be specified by a regular grammar and which contain a fixed number of semantic content fragments.

Future work focuses on understanding utterances, which are still out of a restricted domain, but with an infinite number of semantic content fragments arranged in a tree topology [9].

REFERENCES

[1] J.G. Bauer: *Entwurf und Implementierung eines Systems zum Verstehen von Zeitangaben aus natürlicher, gesprochener Sprache*, master thesis, Institute for Human-Machine-Communication, Munich University of Technology, 1994

[2] A. Hauenstein, E. Marschall: *Methods for Improved Speech Recognition over Telephone Lines*, Proc. IEEE ICASSP 1995 (Detroit, Michigan USA), pp. 425-428

[3] H. Höge: *Statistische Modelle für die Spracherkennung*, Proc. DAGA 1993 (Frankfurt am Main, Germany), pp. 11-30

[4] J. E. Hopcroft, J. D. Ullman: *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, 1979

[5] J. Müller, H. Stahl: *Ein Ansatz zum Verstehen natürlicher, gesprochener Sprache durch hierarchisch strukturierte Hidden-Markov-Modelle*, Proc. KONVENS 1994 (Vienna, Austria), pp. 260-269

[6] H. Ney, D. Mergel, A. Noll, A. Paeseler: *Data Driven Search Organization for Continuous Speech Recognition*, IEEE Transactions on signal processing, vol. 40 (1992), no. 2, pp. 272-281.

[7] H. Ney: *Stochastic Grammars and Pattern Recognition*, Proc. NATO ASI, vol. F75, Springer, 1992, pp. 319-344

[8] J. Piccone: *Continuous Speech Recognition Using Hidden Markov Models*, IEEE ASSP Magazine, July 1990, pp. 26-41

[9] H. Stahl, J. Müller: *A Stochastic Grammar for Isolated Representation of Syntactic and Semantic Knowledge*, Proc. EUROSPEECH 1995 (Madrid, Spain), to be published