Intention-based Probabilistic Phrase Spotting for Speech Understanding

Marc Hofmann and Manfred Lang

Institute for Human-Machine Communication
Technical University of Munich, D-80290 Munich, Germany
{hofmann,lang}@ei.tum.de

ABSTRACT

We present an approach towards probabilistic phrase spotting for evaluating a speech recognizer's utterance hypotheses for inferring the user's intention. The evaluation is done by mapping each word chain on each intention of the intention space. Therefore we create an intention model for each intention as the basis for analysis. As the words of the speech recognizer's utterance hypotheses are assigned confidence levels, we treat these inputs as uncertain observations. We use Bayesian belief networks as mathematical fundament for intention modeling and probability theory for evaluating such word chains. The algorithm considers syntactical and semantical relations between the words within a phrase, evaluating words regarding previously observed words of the current phrase.

1. INTRODUCTION

Naturally and spontaneously spoken utterances for controlling applications are generally a problem for interpreting the speech recognizers´ output as the out-of-vocabulary problem is a massive threat. The application to control is the IT-equipment of an automobile. Pronounced utterances often show omitted endings, expletives which might not be part of the speech recognizer´s vocabulary as well as syntactically wrong formulated sentences. As the number of system functions of such an application is quite restricted, the number of potential utterances to pronounce a goal is restricted as well. In [1] we introduced a system for a syntactic-semantic evaluation of spoken utterances staying abreast of the intention space to cope with this kind of problem. However the utterances to evaluate had been restricted to quite simple structured sentences. In this paper we present an significantly advanced system which is capable of dealing with quite complex utterances. Moreover it proved to be more stable, flexible and robust regarding the out-of-vocabulary problem.

2. BASIC STRUCTURE

The basic idea of our algorithm is mapping the speech recognizer´s *n* best utterance hypotheses on the intention space, respectively on the potential utterances to pronounce the intentions. Fig. 1 pictures the systems´ principle. The user's utterance is the input for a speech recognizer resulting in the *n* best word chains, which are the input vector for our algorithm. The intention library contains all potential intentions a user could have, i.e. all system functions a user can access by speech input. For each intention an intention model is created which is the basis for a syntactic-semantic evaluation. Each utterance hypothesis will be mapped on all intention models regarding the content and syntactic-semantic aspects resulting in an evaluation measure. The utterance/intention combination with the maximum evaluation measure represents the most likely intention.

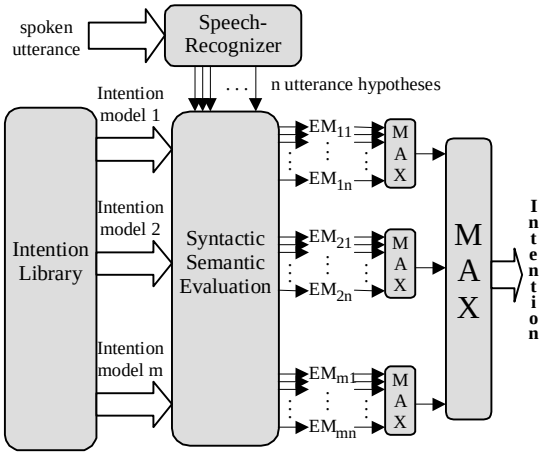


Figure 1: Basic structure of the algorithm

The main components of the system, the intention library and the syntactic-semantic evaluation component will be now described in detail.

3. INTENTION LIBRARY

The intention library covers all intentions a user could have when dealing with our application, i.e. the manipulation of various system parameters, such as changing the volume or playing a special song from CD. For an overview of the structure of an intention model refer to Fig. 2. For syntactic-semantic evaluation of an utterance regarding a special intention, each intention has to be represented by an intention model. Therefore each intention of the library has an utterances space, consisting of a number of potential utterances to

From: ISIMP 2001
IEEE Hong Kong Chapter of Signal Processing

pronounce that intention. For complexity reasons the utterances are created by combining different phrase types / phrases. At word level we refer to phrases as *basicphrases*, that means they don't consist of any further sub phrases, just of words.

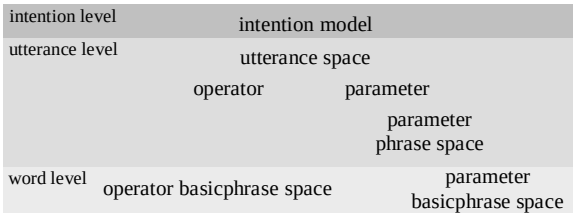


Figure 2: Structure of an intention model

We interpret a typical utterance for controlling an application as a combination of an operator phrase and a parameter phrase. The operator phrase is used to tell the system which system parameter to manipulate, the parameter phrase is used to tell the system the new parameter or the group of parameters, respectively. We use a Bayesian belief network [2] with boolean state variables to model that fact. Fig. 3 shows the topology of that network. We will refer to this network as *intention network*. To put emphasis on the parameter phrase we train the network in that way that a complete observation of the parameter is interpreted as a complete observation of the whole intention. Because the observations of words and phrases are based on the confidence levels of the speech recognizer, i.e. we deal with uncertain observations, the parameter is never completely observed. We use the observation of the operator to support belief in a special parameter phrase. Hence the conditional probabilities are chosen according to the following equation:

$$Intention = (Operator \wedge Parameter) \vee Parameter \quad (1)$$

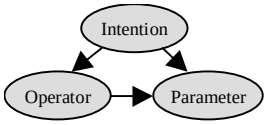


Figure 3: Topology of the intention network

The operator is an operator basicphrase of the operator basicphrase space. To model an operator basicphrase we use the network topology pictured in Fig. 4. Each word node is a boolean state variable to reflect the belief of observing the word which is assigned to that node. An operator basicphrase is only observed completely if all it's words have been observed completely. So the conditional probabilities are trained according to the following equation:

$$operator\ basicphrase = word1 \wedge word2 \wedge \mathrm{K} \quad (2)$$

The parameter phrases are structured and trained by analogy to the operator basicphrases. The word nodes are replaced by nodes which are assigned to the parameter basicphrases. Hence a parameter phrase may consist of more than one basicphrases.

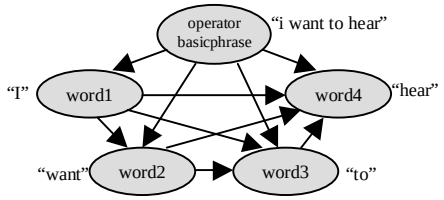


Figure 4: Topology of on operator basicphrase

The parameter basicphrases are modeled in a different way. Fig. 5 shows a typical structure for an parameter basicphrase network. We interpret an parameter basicphrase to be completely observed if its keywords have been observed completely. To the remaining words we refer as optional words. The conditional probabilities of the network of Fig. 5 have to be trained according to the following equations (3),(4),(5):

$$parameter\ basicphrase = (optional \wedge keywords) \vee keywords \quad (3)$$

$$keywords = word1 \wedge word3\,;\ optional = word2 \quad (4), (5)$$

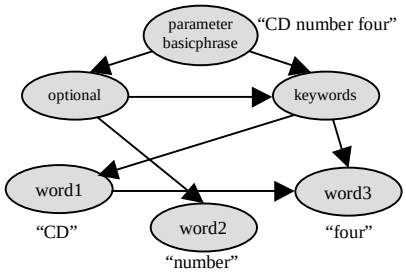


Figure 5: Topology of an parameter basicphrase network

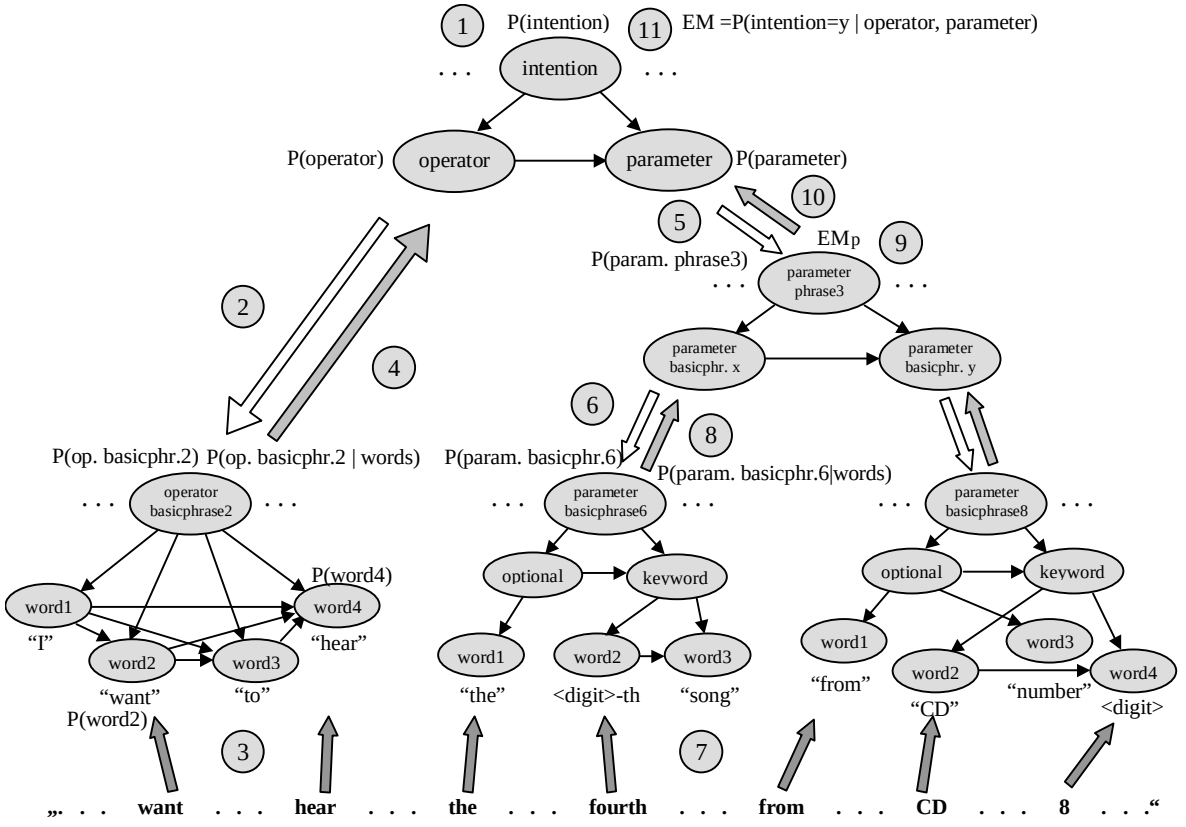
Like the operator on the intention level the optional words are used to support belief in the observation of the keywords. If a word is dealt with as keyword depends on the phrase, generally keywords are the characteristic words of a phrase.

4. SYNTACTIC-SEMANTIC EVALUATION

The algorithm for the syntactic-semantic evaluation of an utterance hypothesis is illustrated in Fig. 6. It considers only one intention. The algorithm will be explained in 11 steps:

① At first the boolean intention node of the intention network is assigned a neutral probability distribution, i.e. both states are equally likely.

② The marginal probability of the operator node of the intention network is calculated and assigned to the root nodes of the operator basicphrase networks to coordinate the basicphrases with intention networks.



original spoken utterance : " well ehmm I want to hear the ehmm fourth song from CD number ehmm 8"

Figure 6: Algorithm for an intention based semantic-semantic evaluation of an utterance hypothesis

③ The utterance hypothesis is parsed for words related to the word nodes of the operator basicphrase network. In Fig. 6 the word "want" has been observed with the uncertainty the speech recognizer quantifies by a confidence level. This observation will be mapped on the network as change in belief, therefore the inference algorithm of Bayesian belief networks has to be extended. Equation (6) enables the inference algorithm to deal with changes in belief. The uncertain observation of a word influences the whole network by the modified joint probability:

$$P_{ob}(root, word\,1, word\,2, ...) = \\ = P(root, word\,1, word\,2, ...) \cdot \frac{P_{ob}(wordx = y)}{P(wordx = y)} \quad (6)$$

At basicphrase level the marginal probability of a word node reflects the assumption that one word of its related vocabulary is part of the utterance hypothesis. A change in belief of a special word will also raise the expectation to observe the remaining words of that phrase, resulting in a stronger weighting of such syntactically and semantically related words. Now the marginal probability as an quantitative measure for the expectation of an observation has to be merged with the uncertain observations by the speech recognizer. Therefore we interpret the range from the marginal probability of a word node $P(wordx=y)$ to 1 as space which is left for observations by the speech recognizer. We map the confidence level on that range and add it to $P(wordx)$, considering the impact of syntax and semantics within a phrase and previously observed words. As the range of the speech recognizer's confidence level c is from 0 to 100 we have to normalize it. Eq. (7) shows the mathematical description:

$$P_{ob}(wordx = y) = \\ = P(wordx = y) + \frac{c}{100}(1 - P(wordx = y)) = \\ = P(wordx = y) + \frac{c}{100} P(wordx = n) \quad (7)$$

The resulting new marginal probability P_{ob} is entered into the basicphrase network as change in belief according to eq. (6). Making use of belief networks' ability to deal with incomplete information, word nodes with unobserved nodes remain not instantiated. That procedure has to be done for all potential operator basicphrases.

④ We determine the operator basicphrase with the maximum a posteriori probability P_{max} of the root node. This is the basicphrase that is described most completely by the utterance hypothesis to evaluate. P_{max} will be assigned to the operator node of the intention network as change in belief, emphasizing

syntactically and semantically related parameter phrases.

⑤ By analogy to step ② the marginal probability of the parameter node will be assigned to the root nodes of all parameter phrase networks.

⑥ By analogy to step ② the marginal probabilities of all parameter nodes of one parameter phrase network will be assigned to the root nodes of all its related parameter basicphrase networks.

⑦ Mapping observed words on parameter basicphrases differs from mapping on operator phrases. At first for each parameter network all potential parameter phrases are created by combining all basicphrases of its basicphrase nodes. This parameter space is the fundament for mapping observations. First the words which are related to the keyword nodes are parsed for, then the optional words are parsed for. Observed words will only be mapped on the basicphrase networks if they fit to the order words occur in the basicphrase. Mapping is done in analogy to step ③.

⑧ For each parameter basicphrase node the basicphrase with maximum root node probability P_{max} is chosen, P_{max} is assigned to the parameter basicphrase node of the parameter network emphasizing the following syntactically and semantically related basicphrases. The parameter basicphrases of the remaining parameter basicphrase nodes have to be treated in the same way. Finally the utterance hypothesis has been mapped on all basicphrase combinations of the current parameter phrase. The combination resulting in the maximum root node probability of the parameter networks is the phrase which is modeled by the utterance hypothesis most completely; this is the combination to choose for that parameter phrase.

⑨ For each parameter phrase model an evaluation measure is calculated to determine the most likely parameter phrase. We determine the percentage *C* of how complete a parameter phrase has been modeled by the utterance hypothesis. As parameter phrases may consist of different numbers of parameters we have to multiply the number of parameters n_p and divide by the maximum number of parameters to be able to compare all parameter phrases. In our system the number of parameters is limited to three parameters. This results in the following equation:

$$C = \frac{P(param = y \mid basicphr\ .) - P(param = y)}{1 - P(param = y)} \cdot \frac{n_p}{3} \quad (8)$$

To calculate an evaluation measure for a special parameter phrase we map *C* on its root node by eq. (9) and eq. (10). The number of observed parameters of a parameter phrase is n_{ip}. In eq. (9) the fraction is used as damping factor to reduce the influence of the syntactic-semantic emphasis described above.

$$\Delta P = (1 - P(param = y \mid basicphr\ .) \cdot C \cdot \sqrt{\frac{n_{ip}}{n_p}} = \quad (9)$$
$$= \frac{1}{3}(P(param = y \mid basicpr\ .) - P(param = y))\sqrt{n_{ip} n_p}$$

And finally we calculate the evaluation measure EM_p of a parameter phrase:

$$EM_p = P(param = y) + \Delta P \quad (10)$$

⑩ The parameter phrase with the maximum evaluation measure EM_p is the phrase of interest. EM_p is assigned to the parameter node of the intention network as change in belief.

⑪ Finally the evaluation measure *EM* for an intention is the a-posteriori probability of the intention node. *EM* reflects how well an utterance to evaluate fits an intention.

5. RESULTS

The algorithm proved to be very robust, recognizing the correct intention even of unclear pronounced utterances and resulting "noisy" utterance hypotheses, i.e. the system is able to cope with the out-of-vocabulary phenomena. The recognition rate for naturally spoken utterances exceeds 90 per cent. Fig. 7 shows the a-posteriori probability, i.e the evaluation measure of all intentions. One intention represents a system function controllable by speech. As there are different modes and parameter types for each system function the real number of intentions tops the number of illustrated intentions by far.

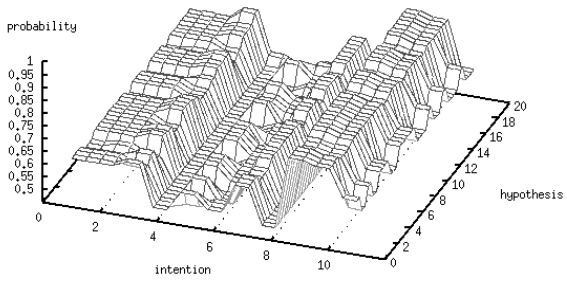


Figure 7: Evaluation measures (probability) of 9 intentions and 20 utterance hypotheses

6. REFERENCES

[1] M. Hofmann, M. Lang, "Belief Networks for a Syntactic and Semantic Analysis of Spoken Utterances for Speech Understanding", ICSLP 2000, Beijing, Vol. II, pp. 875-878

[2] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Morgan Kaufmann, 1988