NAVIGATION IN VIRTUAL WORLDS VIA NATURAL SPEECH

Björn Schuller, Frank Althoff, Gregor McGlaun and Manfred Lang

Institute for Human-Machine-Communication, Technical University of Munich, D-80290 Munich, Germany
(schuller | althoff | mcglaun | lang)@ei.tum.de

ABSTRACT
In this paper we propose a new approach enabling users to intuitively navigate in arbitrary virtual 3D-worlds via natural spontaneous speech. Three different approaches are considered and evaluated: First of all, two stochastic top-down decoders - a one- and a two-pass, split between acoustic and semantic layers. Besides these, a new two-pass decoder is being introduced. Two interfaces allow for multimodal integration: the different decoders can be used as homogenous competing instances already on the syntactic-semantic layer. A context sensitive intention decoder can dynamically constrain their recognition processes and translates their semantic structures into an abstract formal grammar. This concept enables heterogenous connection of additional input modalities on higher levels. The decoder can furthermore provide a measurement for the confidence even of a semantic interpretation based on acoustic confidences and by using rival decoders.

1. INTRODUCTION

Virtual worlds can be used to effectively provide information in different domains like architecture, medical technology, tourist information and education, among others. If we want to give any user intuitive access to this media we have to design interaction as natural as possible. In this respect, we introduce natural speech as a new input modality for the navigation through any arbitrary virtual three-dimensional world.
In a first step, we studied the potential users' acceptance and behavior in a series of Wizard-of-Oz experiments. These initial usability studies clearly show that the recognition of single commands does not satisfy users' needs. But natural speech can highly optimize the workflow if we fulfill two basic requirements: real-time processing and robust recognition coping with out-of-vocabulary words. For a satisfying and effective communication a non-delayed system-feedback is indispensably. The idea to enable a user to navigate in pre-implementation unknown worlds keeps the system tolerant, but experiments showed that nevertheless users tend to apply world-specific vocabulary. This forces us to handle non-integrated words. In this respect, three different approaches have been regarded and evaluated.

2. ADJUSTED NAVIGATION ENVIRONMENT

The virtual worlds are created by an adapted VRML-browser. It can be controlled using different modalities like mouse, keyboard or command shell. An overall defined interface allows connecting future modalities for the navigation. Adding speech as an input modality we introduced temporal and spatial discretisation. For an easier control of the movements in the discrete space we also integrated an incrementation of the step-size. Generally, speech underlies miss-interpretation or -recognition. To enable a user to correct false system actions we implemented an undo-function that naturally should possess a high priority in the interpretation process and can also be activated by a haptic device as fallback solution. A further feature is a repeat function that highly accelerated the workflow. The enhanced functionality claimed for an additional feedback.

3. INTERPRETATING NATURAL SPEECH

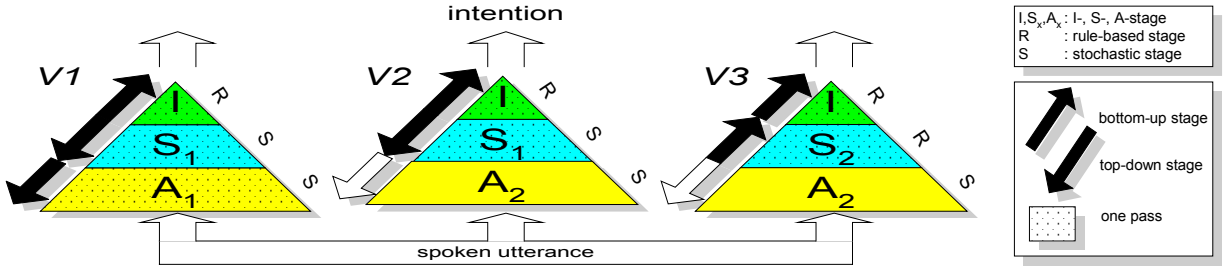

Figure 1: Overview over the three different regarded decoder principles

From: Proc. HCI 2001, Lawrence Erlbaum Ass.

We split the process of interpreting a speech signal into three stages: The acoustic stage (*A*), the semantic-syntactic stage (*S*), and finally the intention stage (*I*). Figure 1 shows the three evaluated realizations. The triangle shapes in the figure stress the reduction of information from the acoustic signal of a spoken utterance over a semantic-syntactic structure to the sheer intention of a user.

3.1 Intention stage

While we make use of two different *A*- and *S*-stages the *I*-stage does not differ in our solutions. It controls an external application via TCP/IP-socket communication, in this case the VRML-browser, by commands obeying an abstract context-free formal grammar. The *I*-stage is realized in a recursive rule-based algorithm, which feed-forward interprets the semantic-syntactic structure delivered by the *S*-stages. It neglects garbage types and integrates contextual and dialogue-historical knowledge in the interpretation. We believe that the a-priori probability for a special user intention P*(I)* should online not be regarded as a static probability as it was seen during training according to usability studies. A system should rather estimate the probability of the occurrence of a user's intention considering the first-order dependencies P*(I|U)* integrating the knowledge of observed habits of the end-user, P*(I|S)* using knowledge of the system model, P*(I|X)* interpreting external influences and P*(I|D)* observing the dialogue history. Following this idea our intention decoder can actively set dynamic priorities for the recognition process according to these parameters to constrain or adjust the vocabulary and the semantic-syntactic models. Due to this fact the different *S*-stages can be in each case regarded as one-pass with the *I*-stage. The idea behind a constant *I*-stage is to allow for pluralistic integration by providing a well-defined interface for the connection with the *S*-stages. Like this, different solutions on lower levels can contribute as competing or complementary instances to the interpretation of the modality user-speech promising more robustness. Furthermore, this principle gives us the opportunity to feed a dialogue instance with a confidence measure on the intention level by comparing coincidences, and enables the system to auto-correct early miss-interpretations by integrating a slower, but more robust engine. Finally the integration allows for work sharing if different aspects of a spoken utterance like the emotional user state should be additionally extracted.

3.2 Semantic-syntactic stages

We considered two different concepts for this stage: A stochastic top-down decoder (S_1) and a rule-based bottom-up decoder (S_2). Their recognition result is a semantic tree structure as shown in figure 2.

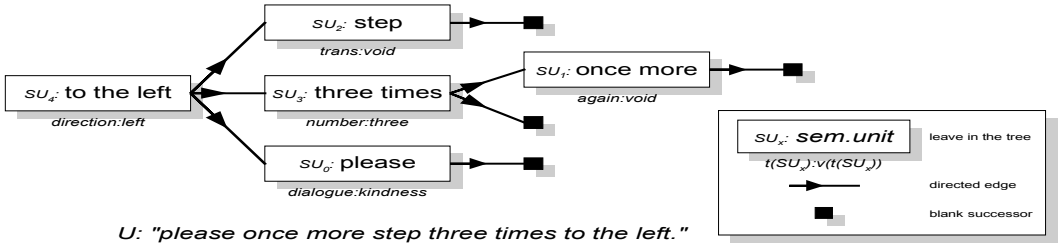

Figure 2: Example of a semantic-syntactic structure

This structure divides an utterance *U* into basic semantic units SU_x by clustering parts of words from single ones up to groups of whole words and assigns these units semantic concepts. Each concept is described by a type $t(SU_x)$ and a value $v(t(SU_x))$, while each type possesses a defined number of successors. These types and their values have to be iteratively designed and sampled in a reference model, which is the common basis of the *S*-stages and the *I*-stage. The semantic-syntactic tree structure has one root unit, connected by directed edges with its successors which themselves are followed by their successors. The direct successors of each unit are arranged in a predefined order according to their type. Each unit is followed by at least an empty successor.

The decoder S_1 is a maximum a-posteriori decoder as introduced in [MS98]. After designing the reference model collected utterances are exemplary assigned to structures by hand via a special designed graphical tool in a first step. In a second step the training can be accomplished semi-automatically with aid of the mentioned tool. With this method we achieve a semantic-syntactic model as basis for the recognition process. This model needs secondary finishing with smoothening techniques to generalize it for unseen occurrences. The root-probability of certain types as well as the whole model can be exchanged online.

The decoder S_2 is a rule-based bottom-up solution. It uses the same reference model as the decoder S_1. Based upon the word chain the assignment of semantic concepts to semantic units is done by lists. We introduce congruent and

conflict types to solve ambiguities in the allocation process by regarding the contextual concepts. Technically we use a score-based system that in analogy to the stochastic processing maximizes the score by counting coincidences and conflicts. Since this method does not necessarily result in a unique solution we defined favored types. The extend of preference can also be set dynamically.

3.3 Acoustic stages

The *A*-stage is either one-pass integrated in the recognition process (A_1) extending the maximum a-posteriori search by the acoustic- and phonetic sub-layers, or delivers the recognized word chain plus optionally single-word confidence measures in an open-microphone manner (A_2) via socket-communication. Providing the single-word confidence measures results in a slower recognition process due to the computing time of the acoustic stabilities. It can therefore be disabled which however negatively influences recognition accuracy.

4. RESULTS

Acceptance studies show that the users are very satisfied with the new modality. The different decoders were tested empirically with 2671 samples and 18 probands at an average age of 26. Our new rule-based approach proves itself to be highly robust: a recognition rate on a word-chain-basis of 96.3% can be achieved with a corpus including 10% out-of-vocabulary utterances. Integrating the acoustic layers we mastered a very fast real-time recognition at 67.7% average recognition rate. Single test persons even achieved a recognition rate higher than 80%. More detailed evaluation results can be seen in the table below showing the quantitative comparison of the system alternatives. The recognition time is given relatively to the average duration of the test utterances. The final reference model consisted of 36 types and overall 62 semantic concepts integrating the different values.

System model V_x	Recognition rate without acoustics	Recognition rate with acoustics	Recognition time
V1: (One-pass) stochastic	86.2%	81.8 %	15.7
		73.6 %	3.8
		52.5 %	2.9
V2: (Two-pass) stochastic		54.4%	1.0
V3: Rule-based	96.3%	72.6%	1.0

Table 1: Evaluation with a test corpus including 10% out-of-vocabulary utterances

5. CONCLUSIONS

Due to the requirements of real time capability and coping with out-of-vocabulary words the one-pass decoder was split between the acoustic and the semantic stages. This results in a loss of recognition rate, since the output of the speech-recognizer may produce syntactically illegal results, which do not correspond with the trained semantic-syntactic model. Nevertheless the gain in recognition-time is amazing. Additionally we can set an acoustic confidence threshold to handle out-of-vocabulary problems in a spotting manner. The rule-based two-pass approach showed itself even more tolerant: it from the basic idea does not forbid any constellation of words. As for portability both methods require a careful design of a reference model. One can use the statistic data achieved in the training for the design of the rule-concept, but an expert could also synthetically set the conflict and congruent types as well as their scores. However expert knowledge is also needed after the stochastic training for the smoothening of the models. As further advantage the rule-set can be adapted faster than performing a retraining to fulfill new requirements like adding extra vocabulary. Both two-pass decoders demanded only a fraction of the memory and system performance compared with the one-pass decoder. These results highly motivate further research in this area. As a next step the system will be enhanced by an integrated intelligent dialog instance. We aim to provide the *n*-best interpretations with single parameter confidence measures on the intention level enabling the dialog instance to selectively solve ambiguities or correct false interpretations by initiating a dialog.

REFERENCES

[MS98] Müller, J. and Stahl, H., *"Speech Understanding and Speech Translation in Various Domains by Maximum a-posteriori Semantic Decoding"*, Proceedings EIS 98, La Laguna, Spain, 1998, Vol. 2 "Neural Networks", pp. 256-267

[AV01] Althoff, F. and Volk, T., *"A Generic User-Interface Framework"*, Proceedings HCII 2001, 9th Int. Conference on HCI, New Orleans, LA, USA, 2001