COLLECTING AND ANALYZING SPOKEN UTTERANCES FOR A SPEECH CONTROLLED APPLICATION

Johannes Müller, Holger Stahl

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

ABSTRACT

To estimate the parameters of stochastic knowledge bases for a speech understanding system, many utterances spoken by many people have to be examined. For the regarded domain of a 'graphic editor', two different manners of collecting training data are discussed. An analysis of spoken commands recorded by a 'Wizard of Oz'-simulation shows that the way to talk to a computer usually depends on how familiar a subject is with a computer.

Keywords: speech understanding, training of stochastic models, 'Wizard of Oz'-simulation

1. INTRODUCTION

A speech understanding system converts a spoken utterance (given as observation sequence O) into its semantic representation (in our approach denoted as semantic structure S) [7]. From the set of all possible S, that S_E has to be found which is most probable given the observation sequence O, i.e. which maximizes the a-posteriori-probability $P(S|O)$. The resulting term can be transformed using the Bayes formula.

$$S_E = \underset{S}{\operatorname{argmax}}\, P(S|O) = \underset{S}{\operatorname{argmax}}\, \frac{P(O|S) \cdot P(S)}{P(O)} \quad (1)$$

Since $P(O)$ is not relevant for maximizing, it can be neglected:

$$S_E = \underset{S}{\operatorname{argmax}}\, [P(O|S) \cdot P(S)] \quad (2)$$

Due to the high variety of S and O it is not possible to estimate $P(O|S)$ directly. Therefore, additional representation levels are necessary. Clearly defined is the word chain W, which can be used to calculate S_E:

$$S_E = \underset{S}{\operatorname{argmax}} \max_{W} [P(O|W) \cdot P(W|S) \cdot P(S)] \quad (3)$$

Eq. (3) can be implemented as 'top-down'-approach for finding that semantic structure S_E, which is the most likely combination of a semantic structure S, a word chain W and the given observation sequence O. We assume statistical independence of all probabilities in the above equations [9] [10] [11].

To solve eq. (3), stochastic knowledge bases (called "models") have to contain the respective probabilities.

- The semantic model delivers the a-priori probability $P(S)$ for the occurrence of a certain semantic structure S.
- The syntactic model delivers the conditional probability $P(W|S)$ for the occurrence of a word chain W given a certain semantic structure S.
- The acoustic-phonetic model delivers the conditional probability $P(O|W)$ for the occurrence of an observation sequence O given a certain word chain W.

The training task is to define the parameters of those stochastic models, this means to estimate the respective probabilities. For the different knowledge bases, different training material is needed: To estimate...

- ... the a-priori probabilities $P(S)$, many semantic structures,
- ... the conditional probabilities $P(W|S)$, many word chains with their corresponding semantic structures,
- ... the conditional probabilities $P(O|W)$, many acoustic realizations (observation sequences) of given word chains

have to be collected and examined.

For the domain 'graphic editor', we collected a lot of command utterances spoken in German language by different subjects, who had to respect following restrictions: Every spoken utterance has to be

- within the domain 'graphic editor',
- without any subordinate clause and
- without any phenomenon of spontaneous speech.[1)]

Since the acoustic-phonetic models can be taken from existing speech recognition systems (e.g. SPICOS [3], SPRING [12]), we only consider the semantic model and the syntactic model. To train the last two knowledge bases, the word chain W, the semantic structure S and the coherence between W and S must be drawn manually for each utterance.

---

1) Examples for phenomena of spontaneous speech: Repetition or omission of words, slips, breaks, stuttering or clearing the throat. Filling words (e.g. "ahm", "hmm", "oh") are treated as insignificant words in the syntactic model.

2. OFFLINE METHOD

This test was done to get a first impression about the way how people talk to a computer. Different subjects should give commands to a speech understanding 'graphic editor' to change a given sequence of pictures either drawn on a sheet of paper or shown on a computer screen [6]. Every picture has two frames titled "window before" and "window afterwards". The frames represent the windows of the 'graphic editor' on the monitor including coloured two-dimensional objects.

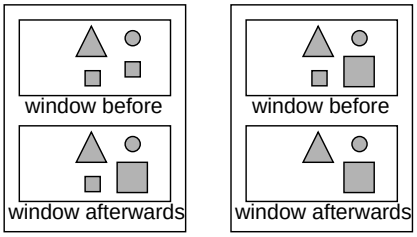

Figure 1: Two examples of pictures given to subjects

It could be observed that (independent from the technical background) many subjects tend to use a computerized language in front of a computer screen whereas they speak longer sentences if they are shown a sheet of paper.

If the pictures are drawn on a sheet of paper, the 'offline method' does not need any computer, but only a tape recorder for storing the spoken utterances. Because of that, this experiment can be undertaken nearly everywhere and is especially suitable for people, who are afraid of talking to a machine.

The great disadvantage is the strongly limited variety of the utterances (limited choice of forms, colours, sizes etc.) because the pictures (i.e. the respective command intentions represented by the semantic structure *S*) are fixed. Hence, it is not possible to estimate semantic parameters, so the offline method is not a suitable data acquisition to train the semantic model.

Nevertheless, it is possible to estimate syntactic parameters, in our case the probabilities $P(W|S)$.

3. 'WIZARD OF OZ'-SIMULATION

To simulate a real application with an unlimited choice of semantic contents within the 'graphic editor' domain, we designed a 'Wizard of Oz'-simulation (WOZ). Within a WOZ, a human 'wizard' simulates the speech understanding part of the computer, while the subject believes to work on a real application. (WOZs are described e.g. in [1], [2] or [4].) The subject's task is to originate any graphics on the screen only by speaking commands. Objects like cones, spheres, cuboids or cylinders can be created, altered or deleted. Since all the subjects think really to work with a computer application, a WOZ seems to be the only way for collecting authentic speech data.

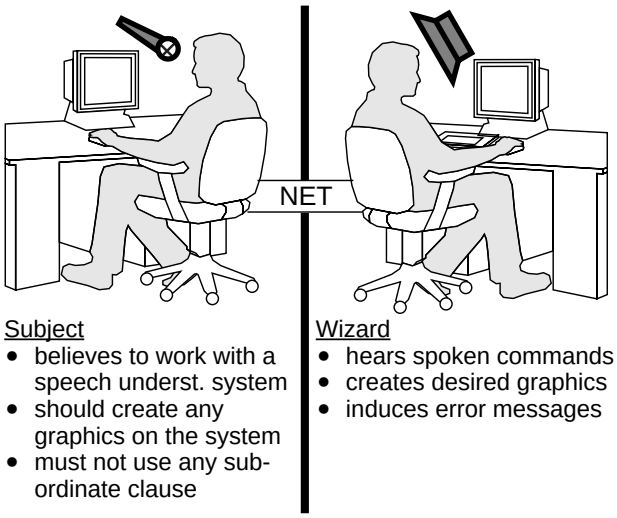

Figure 2: Principle of the "Wizard of Oz"-simulation

With the WOZ-experiment, 1915 German utterances from 33 different speakers were collected. 1843 utterances are valid (i.e. within the domain 'graphic editor', without any subordinate clause, no phenomena of spontaneous speech). The resulting vocabulary has 854 word entries.

4. QUANTITATIVE UTTERANCE SIZES

For the design of a speech understanding system, it may be important to know something about information quantity, complexity, size and duration of the utterances. Within our statistical approach, we are able to extract quantitative data from the observation sequence *O* (representing the speech signal), its word chain *W* and its corresponding semantic structure *S*.

- An observation sequence can be characterized by its time duration T_O .
- A word chain can be characterized by the number of significant and insignificant words.
- A semantic structure[1)] [7] [8] can be characterized by following values: The number N of semuns (which is equal to the number of significant words [11]) stands for the quantity of the given information. The nesting depth D (i.e. the number of se-

[1)] A semantic structure *S* is a tree consisting of a finite number N of semantic units (simply called 'semuns') s_n :

$$S = \{s_1, s_2, \ldots, s_N\} .$$

Each semun $s_n \in S$ with $1 \le n \le N$ is an (X+2)-tupel of a type $t[s_n]$, a value $v[s_n]$ and $X \ge 1$ successor-semuns $q_1[s_n], \ldots, q_X[s_n] \in \{s_2, \ldots, s_N, \text{blnk}\} \setminus \{s_n\}$:

$$s_n = (t[s_n], v[s_n], q_1[s_n], \ldots, q_X[s_n]) .$$

(We are currently using semuns with $1 \le X \le 5$ successors.) The semun s_1 is defined as the root of *S*. Each other semun $s_2, \ldots, s_N$ is marked exactly once as a successor semun. The special semun 'blnk' has the type *t*[blnk] = blnk, no value and no successor.

muns along the longest path within *S*) shows how detailed the information is.

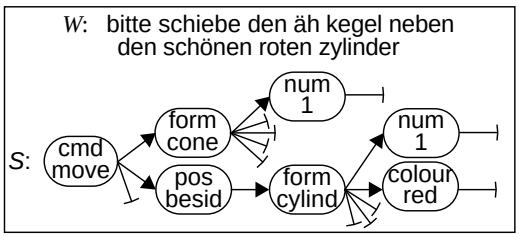

Figure 3: Word chain *W* and semantic structure *S*

As an example, fig. 3 shows the word chain *W* and semantic structure *S* of the German utterance "bitte schiebe den äh kegel neben den schönen roten zylinder" ("please move the ahm cone beside the pretty red cylinder") delivering the following values:

- The total number of words is 10.
- The number of insignificant words is 3: "bitte" ("please"), "äh" ("ahm") and "schönen" ("pretty").
- The number of significant words equal to the number *N* of semuns is 7.
- The nesting depth *D* is 4.

5. ANALYZING THE UTTERANCES

For the WOZ-experiment, all subjects – male (m) or female (f) – are asked how often they use a computer: frequently (i), occasionally (o) or never (n). Taking the observation sequences, word chains and semantic structures of all the WOZ-utterances, the following observations can be done distinguishing these groups:

Fig. 4 shows the average duration $\overline{T_O}$ of the spoken commands. It can be observed that (m)-subjects speak shorter utterances than (f)-subjects and the duration is rising from (i)- to (n)-subjects. Notice that (n)-subjects ($\overline{T_O}$ = 4, 92 s) need on average 1.5 times longer to speak a command than (i)-subjects ($\overline{T_O}$ = 3, 27 s).

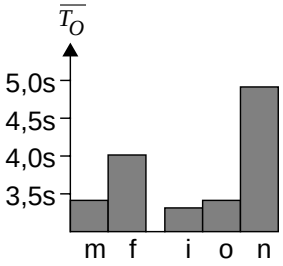

Figure 4: Average duration $\overline{T_O}$ per utterance

An utterance counts as valid, if it is within the domain 'graphic editor' and does not contain any subordinate clause or phenomena of spontaneous speech. In fig. 5, it seems that (f)- and (o)-subjects like to check the abilities of that system, whereas (m)-, (i)- and (n)-subjects tend to respect the given restrictions.

Sometimes it could be observed, that some (f)-, (o)- or (n)-subjects speak comments ("I feel bored", "hmm, what should I do" or "oh, you are a perfect computer").

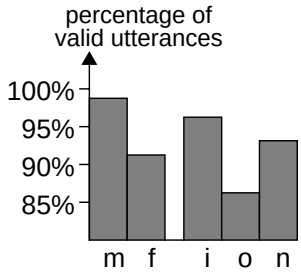

Figure 5: Percentage of valid utterances

We observed that (i)-subjects, who are familiar with computers and know that the usual human-machine-dialogue is inflexible due to strict rules, use computer-like formulations with very short commands, a minimum of significant words and very few insignificant words. On the other hand, (o)- and (n)-subjects utilize more insignificant words within longer commands (which are sometimes quite similar to a human-human-dialogue). Fig. 6 shows that in comparison to (i)-subjects, (o)- and (n)-subjects give commands containing more words with more insignificant words. A similar statement can be done comparing male and female subjects.

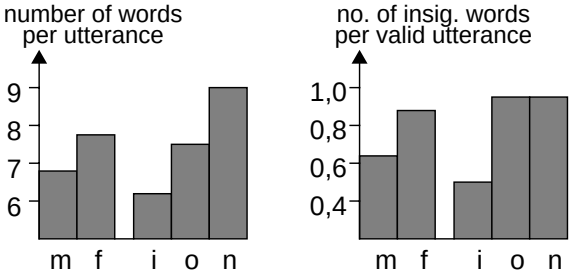

Figure 6: Average number of words and average number of insignificant words

Fig. 7 displays the complexity and the details of the semantic content falling from (n)- to (i)-subjects. The first group usually knows the problems of speech understanding and tries to avoid semantic complexity. The latter group is quite naive on what the computer is able to do and supposes that everything a human can decode, also a computer can decode.

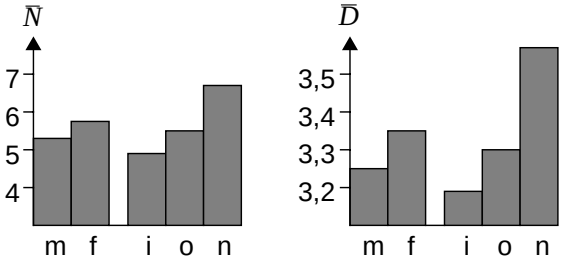

Figure 7: Average number $\overline{N}$ of semuns and average nesting depth $\overline{D}$ per valid utterance

It may be important to know some occurring maximum sizes observed in all collected spoken utterances:

- The longest duration of a spoken utterance is $T_{O,\max} = 16,9\ \mathrm{s}$.
- The longest word chain contains 25 words.
- The largest amount of insignificant words in a valid word chain is 5.
- The largest amount of significant words in a valid word chain equal to the maximum number of semuns in a semantic structure is $N_{\max} = 18$.
- The maximum nesting depth of a semantic structure is $D_{\max} = 6$.

6. CONCLUSION

A speech understanding system must consider the users working on it. The habit of using the computer strongly correlates to the manner of talking to a speech understanding system. Complexity, size, duration and vocabulary of spoken utterances and the corresponding information content may differ in a wide range, but often depend on the user himself. Examinations show, that a suitable dialogue could unconsciously limit these varieties by giving the user an idea how to express his actual intention [5] [13].

Our future research will focus on designing dialogue models in order to get the user to say what the system can understand. Furthermore, situation dependent models can improve the robustness of the system.

REFERENCES

[1] M. Blomberg et al.: *An Experimental Dialogue System: Waxholm*, Proc. Eurospeech 1993 (Berlin, Germany), pp. 1867-1870

[2] H. Dybkjaer, N.O. Bernsen, L. Dybkjaer: *Wizard-of-Oz and the Trade-off between Naturalness and Recognizer Constraints*, Proc. Eurospeech 1993 (Berlin, Germany), pp. 947-950

[3] H. Höge: *SPICOS II - a Speech Understanding Dialogue System*, Proc. ICSLP 1990 (Kobe, Japan), pp. 1313-1316

[4] I. Katunobu et al.: *Collecting and Analyzing Nonverbal Elements for Maintenance of Dialog Using a Wizard of Oz Simulation*, Proc. ICSLP 1994 (Yokohama, Japan), pp. 907-910

[5] D. Luzzati, F. Neel: *Dialogue Bahvior Induced by the Machine*, Proc. Eurospeech 1989 (Paris, France), vol. 2, pp. 601-604

[6] A. Marx: *Generieren und Optimieren von Sprachmodellen, sowie regelbasiertes Textverstehen am Beispiel eines sprachgesteuerten Grafik-Editors*, diploma thesis, Institute for Human-Machine-Communication, Munich University of Technology, 1994

[7] 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

[8] J. Müller, H. Stahl: *Die semantische Gliederung als adäquate semantische Repräsentationsebene für einen sprachverstehenden 'Grafikeditor'*, Proc. GLDV-Jahrestagung 1995 (Regensburg, Germany), to be published

[9] R. Pieraccini, E. Levin: *Stochastic Representation of Semantic Structure for Speech Understanding*, Proc. Eurospeech 1991 (Genova, Italy), pp. 383-386

[10] R. Pieraccini, E. Levin, E. Vidal: *Learning how to Understand Language*, Proc. Eurospeech 1993 (Berlin, Germany), pp. 1407-1412

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

[12] K. Wothke et al.: *The SPRING Speech Recognition System for German*, Proc. Eurospeech 1989 (Paris, France), vol. 2, pp. 9-12

[13] E. Zoltan-Ford: *How to Get People to Say and Type what Computers Can Understand*, International Journal of Man-Machine Studies, vol. 34 (1991), no. 4, pp. 527-547