A SYSTEM FOR IMPROVING THE RECOGNITION OF FLUENTLY SPOKEN GERMAN SPEECH

Joachim Mudler

Institut fur Nachrichtentechnik, Technische Universitat Braunschweig
Schleinitzstr. 23, D-3300 Braunschweig, Fed. Rep. Germany

ABSTRACT

A research project for improving the recognition of fluently spoken German speech is presented. The work is in progress at present.*

It should be investigated, how far aspects of semantics and inferences could improve the automatic speech recognition. The work is part of a speech recognition system that receives speech signals, converts them into forms suitable for further actions and finally puts out the spoken text in characters. The system itself operates at three stages. Within the first one the signal analysis is performed using a well-known method. This analysis segments the signal into certain subword units and, for each segment, produces a set of weighted candidates. At the second stage these candidates are used to generate weighted word hypotheses with the aid of an extensive lexicon. The hypotheses have to be verified or falsified within the following processing steps at the third stage. Thereby the algorithm uses a best-first strategy (hypotheses with highest weight first).

Besides syntactic/grammatical aspects, semantic analysis and inferences mentioned above are the methods, that should lead to a certain text-comprehension.

I INTRODUCTION

This paper presents the basic concept of a speech recognition system for fluently spoken German speech. The goal is to improve recognition using methods from the field of Artifical Intelligence.

The system operates at three stages (fig. 1). The first one performes the signal analysis. The speech signals are segmented into certain subword units and, for each segment, a set of weighted candidates is produced. This is described in chapter II. The second stage uses these candidates to generate weighted word hypotheses. Chapter III explains how the word hypotheses are produced. The third stage produces a set of weighted sentence hypotheses and performs the recognition. The algorithm is described in chapter IV.

An extensive lexicon is necessary to execute the last two stages. The chapters III .A and IV.A describe the lexicon entries.

From the wave form of the speech signal we pick out a section which we call the signal window or window. At present, we presuppose that the size of the window exactly corresponds to one spoken sentence.

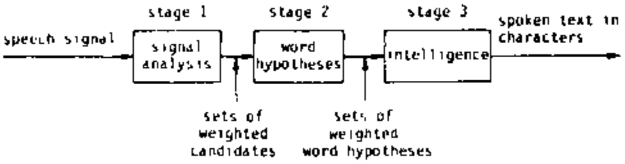
stage 1
stage 2
stage 3
speech signal
signal analysis
word hypotheses
intelligence
spoken text in characters
sets of weighted candidates
sets of weighted word hypotheses

figure 1: system overview

II SIGNAL ANALYSIS

The speech signals received by the speech recognition system have to be converted into forms suitable for further actions. The signal analysis uses a well-known method (Dettweiler, 1981; Ruske and Schotola, 1978).

The analysis segments the signal into certain subword units, socalled demisyllables. A syllable contains a vocalic nucleus with an initial conconant cluster preceding and a final consonant cluster following the vowel. Now a demisyllable is defined as one or the other part of a syllable, if the syllable is cut into two parts somewhere during the vowel.

Regarding this method, there exists an inventory of about 1300 demisyllables for unrestricted German speech, if all initial consonant clusters are combined with all possible vowels (about 650), and all possible vowels are combined with all final consonant clusters (about 650, too).

By means of this method the speech signal is segmented into demisyllables, and for each segment, a set of weighted candidates is produced. For demonstration we presuppose that this set of weighted candidates is a complete one.

So, at the end of the analysis, a chain of complete sets of weighted candidates for each

* This work is supported by the Deutsche Forschungsgemeinschaft (DFG).

possible position is given, that can be used for the next stage of the recognition system (fig. 2).

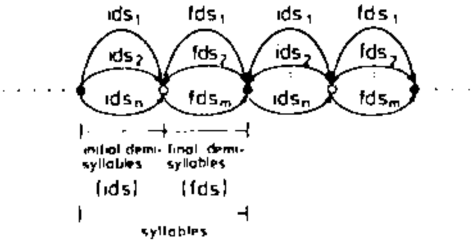

figure 2: a chain of demisyllables and sets of weighted candidates

III WORD HYPOTHESES

A. Lexicon Entries

Within the lexicon there exist some entries that are used for generating word hypotheses, regarding the weighted demisyllable segmentation of the preceding stage. For each word in the lexicon, its phonetic transcription using demisyllables and its number of syllables are listed.

B. Generation of Word Hypotheses

To generate word hypotheses the lexicon is passed through successively word by word. Taking one word out of the lexicon its weight is calculated for each possible position within the chain of demisyllables using the weights of the respective subword units.

If we have N syllables within the chain (that means 2N demisyllables), a word with the length of one syllable could appear at N positions, a two-syllable word at (N-l) positions and so on. Having processed all words of the lexicon, this method leads to a complete set of word hypotheses for the contents of the respective window (fig. 3).

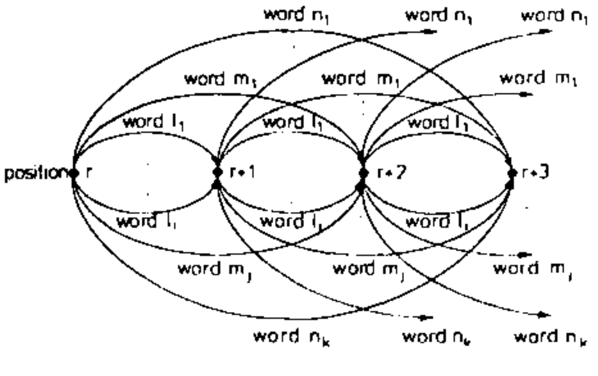

figure 3: word hypotheses

IV INTELLIGENCE

A. Lexicon Entries

Besides the entries mentioned above, the lexicon also contains syntactic/grammatical and sematic informations for each word. The words are classified into categories, e.g. noun, adjective, preposition, verb, auxiliary verb. The inflectable words have a set of possible grammatical forms as an entry, and, perhaps, a semantic marker description. When a certain word hypothesis is processed, the information taken from the lexicon is used to answer pending requests or to produce inferences concerning other possibly appearing words or structures within a sentence.

B. Algorithm

As regards the weighted word hypotheses, now the purpose is to attain correct and meaningful sentences by finding pathes from the beginning to the end of the section of the window.

To start with the algorithm the best word hypothesis is selected. The respective lexicon entries of this word (the structure hypotheses and their requests) are transferred to a request file which is part of the intelligence stage as shown in figure 4. The structure of the request file is tree-oriented. Within this tree pathes must be traced that could lead to the acceptance or rejection of the actual hypothesis.

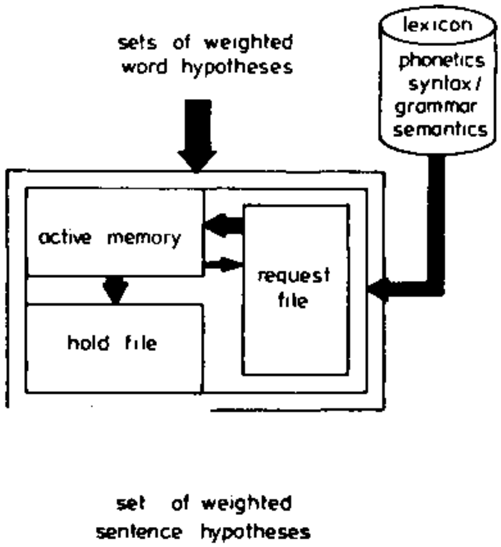

figure 4: intelligence stage

In addition, the length of the window and the (relative) positions of word hypotheses are always known. In the following the recognition algorithm performs an expectation-based analysis (Riesbeck and Schank, 1978).

Supposing the selected word as a noun the request file could have an entry like:

Can an adjective precede the noun directly?

If there could be at least one adjective the path leads us to requests like:

Are the adjective and the noun grammatically congruent?
and
Are the adjective and the noun semantically congruent?

A request as an inference from the lexicon could look at the pending word hypotheses or at a hold list, where recognized parts of a sentence, e.g. noun phrases, are listed. In the first case the request generally concerns word categories. Such a request could look like:

Which words with category cat could precede (or follow) the actual hypothesis up to position <i> (or starting from position <j>)?
(The actual hypothesis starts at position <i> and ends at position <j>).

If the request could be satisfied the next requests would concern the grammatical and semantical congruences, taking the answer with the best weight first. If any request is rejected the path that led to this request is removed from the request file up to that point where an alternative path can be traced. At all points, answers to a request are only regarded, if they have a certain minimal weight in order to limit the number of the following requests. Whenever a word hypothesis or a structure hypothesis can be verified in the described way that fact leads to a new selection of the still pending hypotheses. If any part of the section within the window had been verified starting at any word hypothesis, and if there are still pending hypotheses, respectively - with other words - so far there exist only parts of a way from the beginning to the end of the window, then the actually best weigthed word hypothesis has to be considered next. The processing ends if there exists at least one path through the word hypotheses that is accepted. Generally, there would be more than one. Each path has a weight, that is calculated from the word weights (fig. 5).

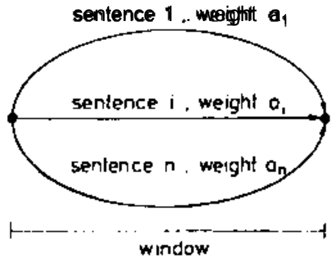
sentence 1 , weight a_1
sentence i , weight a_i
sentence n , weight a_n
window

figure 5: sentence hypotheses

If no path could be found because of the supposed minimal weigth, this threshold has to be reduced and a new trial has to start.

So, finally there had been a transformation from sets of weighted word hypotheses to a set - an essentially smaller one - of possible sentences. The recognition algorithm takes that sentence for granted that has the best weigth. The spoken text can be printed in characters.

V CONCLUSION

In this paper an overview of an automatic speech recognition system is presented. The system receives speech signals and performs the recognition of the fluently spoken German texts.

Some remarks should complete the description of the recognition system. They show what aspects should be investigated in addition.

It is possible to limit the sets of word hypotheses by taking only those words that exceed a certain threshold. For this case it becomes necessary to have a feed-back to the subword level, because it is possible that only such a word can satisfy a path that did not exceed the threshold and that had not been considered within the hypotheses so far.

It is planned not to limit the window size. So, the window could also contain only a part of a sentence, parts of more than one sentence or more than one sentence.

There are a lot of methods to realize the weighting of demisyllables, words and sentences.

Because of an internal knowledge representation that could be used to handle inferences a question-answering system could be attached to the system. The representation follows the ideas of Schanks conceptual dependency theory (Schank, 1975).

REFERENCES

[1] Dettweiler, H., "An Approach To Demisyllable Speech Synthesis of German Words". In Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, Atlanta, Georgia, USA, March 1981, pp. 110-113.

[2] Ruske, G. and Schotola, T., "An Approach to Speech Recognition Using Syllabic Decision Units". In Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, Tulsa, Oklahoma, USA, April 1978, pp. 722-725.

[3] Riesbeck, C.K. and Schank, R.C., "Comprehension by Computer: Expextation-based Analysis of Sentences in Context". In Levelt, W.J. and Flores Dervais (eds.), "Studies in the Perception of Language", John Wiley and Sons, 1978, pp. 247-293.

14] Schank, R.C., "Conceptual Information Processing", Amsterdam, North-Holland Publ. Co., 1975.