



The SPRING Speech Recognition System for German

K. Wothke

U. Bandara, J. Kempf, E. Keppel, K. Mohr, G. Walch

IBM Scientific Center Heidelberg

ABSTRACT

An experimental speech recognition system was developed for German using an already existing technology reported elsewhere (1). The system recognizes complete sentences when the words are spoken with a small pause in between. The user has to train the system in advance by uttering 110 short sentences. The size of the system's vocabulary is presently limited to about 1300 words, with a coverage being 58% of the running words in a newspaper text from the commercial discourse domain. The system uses 60 allophones and a statistical trigram model made out of a text corpus of 14 million words. The recognition accuracy is over 95%.

1 INTRODUCTION

Since 1988 a research group at the Scientific Center/Heidelberg of IBM is developing a speech recognition system for German. The project is called SPRING (= SPEech Recognition IN German).

The methodological and the technical basis for the German system is the English speech recognition system TANGORA developed by F. Jelinek et al. at IBM Thomas J. Watson Research Center in Yorktown Heights, USA (1). This is a speaker dependent recognizer for sentences uttered in isolated-word mode. It is able to recognize a vocabulary of about 20000 words. (Each inflected word form is counted as one word.)

The main aims of the SPRING project are:

1. Development of a research prototype of a large vocabulary speech recognition system for German on the basis of the TANGORA technology.
2. Development of a research prototype of a voice activated typewriter for German.
3. Identification and solution of speech recognition problems specific to German.

At present there exists an experimental German recognizer for about 1300 words. It is the subject matter of this paper. - Chapter 2 contains a short explanation of fundamental theoretical principles of the speech recognition procedure. This is followed by a description of the architecture of the system developed on this basis. Chapter 4 describes the current state of the German specific components.

2 THEORETICAL PRINCIPLES

The TANGORA speech recognition system is not a knowledge based system.

Instead it is based on a statistical approach proposed by Jelinek et al. (1).

Let $W = w_1, \dots, w_N$ be a sequence of N words, and $A = a_1, \dots, a_M$ be a string of acoustic information extracted from the speech signal. The task of the system is to find the word sequence W s which maximizes the probability that the words W were uttered, given that A was observed, $P(W|A)$:

$$P(W_s|A) = \max_W P(W|A) \quad (1)$$

Application of Bayes' rule to the left hand side yields

$$\max_W P(W|A) = \max_W \frac{P(A|W) \times P(W)}{P(A)} \quad (2)$$

where

- $P(A|W)$ is the probability that the acoustic information A will be extracted from the speech signal, if the word sequence W is uttered.
- $P(W)$ is the probability that the word sequence W occurs in the language.
- $P(A)$ is the probability that the acoustic string A occurs. Since it is independent of W it can be treated as a constant in the maximizing process.

3 SYSTEM ARCHITECTURE

The system architecture is shown in figure 1. The four steps of the recognition process are shown in the right part of the figure. The tasks of these steps can be pointed out in terms of equation 2.

- The **signal processor** has to extract a sequence of so-called acoustic labels $A = a_1, \dots, a_M$ from the speech signal. (Computation of A).
- The tasks of the **fast acoustic match** and of the **detailed acoustic match** are to find those words which make a high contribution to the term $P(A|W)$, i.e. to find those words which are most likely to produce the observed label sequence A.
- The **language model** computes for a sequence of words $W = w_1, \dots, w_N$ the probability of occurrence in the language. (Computation of $P(W)$).

The four recognition steps will now be explained in more detail.

The whole system is implemented in a PC/AT with up to four special cards carrying fast signal processors and fast memories.

Signal Processor

Outside the AT the speech signal is amplified and digitized (20K samples/sec., 12 bits/sample). The data are then read into a buffer on a signal processor card. Every 10 ms an FFT is performed over a window of 25.6 ms. From the resulting frequency spectrum a vector of 20 elements is taken. Each element represents the energy density of one out of 20 frequency bands between 200 Hz and 8 kHz. The spectrum is adjusted to account for an ear model.

Each vector is then compared with 200 speaker dependent prototype vectors. The identification number, which is called acoustic label, of the most similar prototype vector is taken and sent to the subsequent processing stages. Each acoustic label corresponds to a speech signal frame of 25.6 ms duration taken from the whole signal every 10 ms. A label needs 1 byte to be represented.

By this process the data rate is reduced from 30000 to 100 bytes/sec.

The speaker dependent prototype vectors are generated from the language specific prototype vectors during a training of the system with a speech sample of 5 minutes.

Fast Acoustic Match

The fast acoustic match determines for every word of the reference vocabulary the probability with

which it would have produced the sequence of acoustic labels observed from the speech signal. The probability of a word is calculated until either the end of the word is reached or the probability drops below a prespecified level.

The fast match uses the following reference units for the determination of this probability:

- a phonetic transcription for each word in the reference vocabulary, including relevant pronunciation variants.
- a hidden Markov model (HMM) for each allophon used in the phonetic transcription.

The phonetic transcriptions were generated automatically using a set of phoneticization rules. These transcriptions were then revised manually. Missing pronunciation variants were added.

The HMM of an allophon describes the probability with which a substring of the sequence of acoustic labels corresponds to the allophon. The Markov models are language specific and the output and transition probabilities are trained to individual speakers.

The Markov model of the phonetic transcription of a word is the chain of the Markov models of its allophones.

Usually the fast match finds 100-300 word candidates with sufficient probability for a given sequence of acoustic labels.

Language Model

The language model receives from the fast acoustic match a set of word candidates. For each of these candidates it determines the probability with which it follows the words which have already been recognized.

For this process the language model uses probabilities of single words, word pairs, and word triples. These probabilities have been estimated for all words in the vocabulary using large text corpora.

The word candidates with the highest combined probabilities supplied by the fast match and the language model are selected and passed to the detailed match, usually 10-30 in number.

Detailed Acoustic Match

The detailed acoustic match computes for each word received from the language model the probability with which it would produce the acoustic label sequence observed by the signal processor.

The main differences to the fast acoustic match are that the detailed match does not perform this

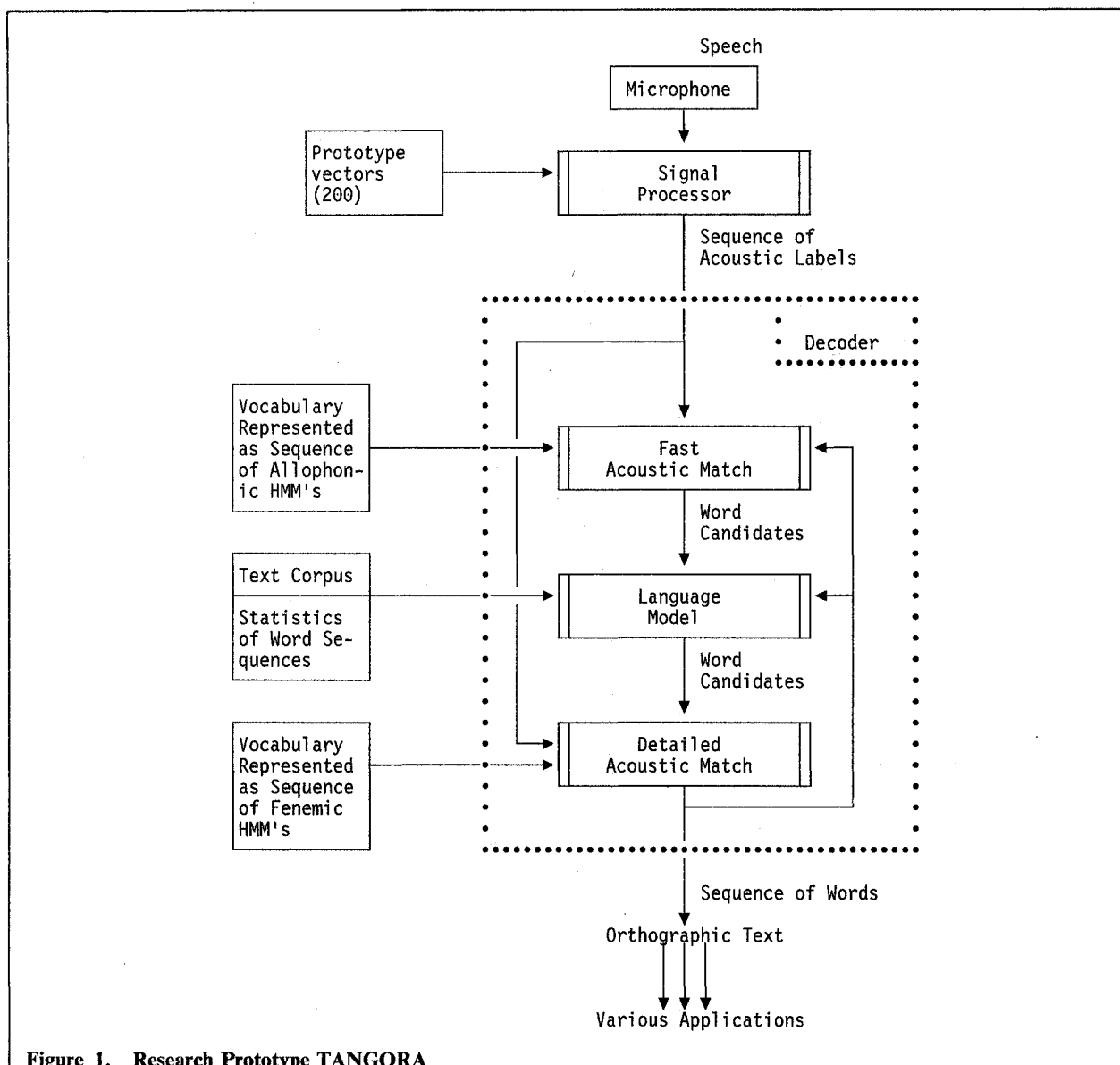


Figure 1. Research Prototype TANGORA

process for all words of the reference vocabulary but only for those received from the language model, and that it does not use phonetic transcriptions and HMM's of allophones as reference units. Instead, the detailed match uses HMM's of so-called fenemic phones. A fenemic phone is an artificial sound unit which usually corresponds to one acoustic label. In the system it is represented by an HMM. The number of fenemic phones is equal to the number of different acoustic labels. The HMM's of the words are the concatenation of the fenemic phones. The HMM's are again language specific. They are constructed automatically from the acoustic data of several speakers. Later their output and transition probabilities are adapted to individual users of the system by an analysis of their acoustic data.

The detailed match determines the most probable words received from the language model.

The three probabilities of the fast match ($P(A|W)$), of the language model ($P(W)$), and of the detailed match ($P(A|W)$) are then combined for the most likely sequences. At the end of each hypothesis the fast match, the language model, and the detailed match are started again.

4 LINGUISTIC COVERAGE

The current German system is a research prototype with a very limited linguistic coverage of German. It was mainly developed in order to learn how to adapt the English TANGORA to German.

Both the hardware and the software of the English system is independent from the language specific

data. Therefore, for the adaptation to German only the language specific data had to be provided. Following is a short description of the German system in terms of these data.

The 1300 words currently recognized by the system are the words with the highest frequency in the financial and commercial sections of the newspaper *Mannheimer Morgen*. For this vocabulary about 3000 phonetic transcriptions of the most relevant pronunciations were semi-automatically generated. With statistical methods those transcriptions were extracted from this initial set which were actually uttered by our test speakers. The resulting final set of about 2200 transcriptions is used by the fast match.

For the phonetic transcriptions we defined a set of 60 allophones which contains

- 25 vowels,
- 30 consonants,
- 3 system specific phones (NULL phone, pause, sentence boundary)
- 2 offglides.

For each of these allophones an HMM was allocated, which is needed by the fast match. The initial statistics for these HMM's presupposes the generation of the German prototype vectors. The HMM's and the prototype vectors were computed in the following way:

1. Recording of both a German and an English training text, which contain the language specific allophones with sufficient frequency. The training text was spoken by seven German speakers.
2. After the processing of the German speech signal the 200 system prototypes for German were computed. These prototypes become also the basis for the construction of the HMM's used by the detailed match.
3. English user training for all of these speakers. By this process we generated an HMM statistics for each English allophon with respect to the German prototype vectors. - The training is performed with a forward/backward algorithm (2).
4. If a similar English allophon existed for a German allophon, then we took the HMM of the English allophon as that of the German allophon.
5. If there was no similar English allophon we combined HMM's of several English allophones in order to create an HMM for the German allophon.

For the language model we estimated a frequency statistics for individual words, word pairs and word triples of the vocabulary. The text base for this statistics is a large text corpus of newspaper articles in the commercial discourse domain.

The HMM's of both fenemic phones and of words, used by the detailed acoustic match, were generated automatically from recordings of utterances of each word from the reference vocabulary. For this task we composed a set of sentences which covers the whole vocabulary. The sentences were spoken by seven speakers.

We did not instruct these speakers to confine themselves to a specific kind of pronunciation. In order to ensure that the fast match does not use too many transcriptions but only those of the pronunciation variants actually uttered by our speakers and used by the detailed match, we extracted these transcriptions with the aid of a Viterbi algorithm (2).

The HMM's of words, used by the detailed match, were constructed by averaging those utterances of a word, which were assigned to the same phonetic transcription with the aid of the Viterbi algorithm.

5 CONCLUSIONS

The SPRING speech recognition system for German is developed on the basis of the English system TANGORA. The development of the current prototype was possible without any changes in the hardware and software of the English system. Only the language specific data had to be provided. They comprise the German prototype vectors, reference vocabulary in orthographic, phonetic and fenemic representation, inventory of allophones, language model statistics and HMM's of allophones and fenemic phones. Furthermore, the prototype vectors and HMM's of allophones and fenemic phones had to be trained for each speaker. (The limited size of this paper does not admit a description of the training procedures).

The recognition accuracy of the German system is over 95%.

Our present aim is to develop a recognizer for 10000 words within 1989 and to solve problems specific to German.

REFERENCES

- 1 Jelinek, F. (1985): The development of an experimental discrete dictation recognizer. In: Proceedings of IEEE 73(1985)11. PP. 1616-1624.
- 2 Rabiner, L. R./Juang, B. H. (1986): An introduction to Hidden Markov Models. In: IEEE ASSP Magazine. January 1986. PP. 4-16.