2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

VOICE CONTROLLED MAIL ORDERING VIA TELEPHONE USING SPREIN

H.W. Ruehl

Philips Kommunikations Industrie AG, Thurn-und-Taxis-Str. 14, D-8500 Nuernberg, FRG

Abstract

For a mail order application, the SPREIN voice control system has been developed and tested in a field trial. It consists of a dialogue and communications controller and several telephone conversation units (TCU). Each TCU contains a speaker independent isolated word recognizer, an unlimited vocabulary speech synthesizer and a telephone line interface. Ordering is done fully automatically using synthetic speech for user guidance and feedback.

The device has been developed in 1986 to 1988, and has been evaluated by the German Telekom in 1988 to 1990. In off-line tests, the recognizer's average error rates improved from 2.7% in 1988 to 1.8% in 1990. A field trial with voluntary users revealed that the main weakness of the system is its slow speed of data entry compared to keyboard entry or human order acceptance via telephone. In spite of its drawbacks, users accepted the system and wanted it to to stay on-line after the end of the field trial.

1. Introduction

To evaluate voice I/O technology for use in public telephone networks, the German telecommunication company Deutsche Bundespost TELEKOM started the project SPREIN (= SPRach-EINgabe) in 1985. It was intended to design a fully automatic mail ordering assistance based on speaker independent recognition of isolated words, and speech synthesis.

The system was supposed to fulfill several purposes:
- provide a demontration and a prototype for a new kind of data entry via telephone
- help evaluate state of the art for voice I/O technology for the German language,
- prove usefulness and user acceptance of speech technology for the public telephone network.

SPREIN is intended to be located at public exchanges, easily configurable to new applications, usable by commercial customers to add voice data entry to their applications.

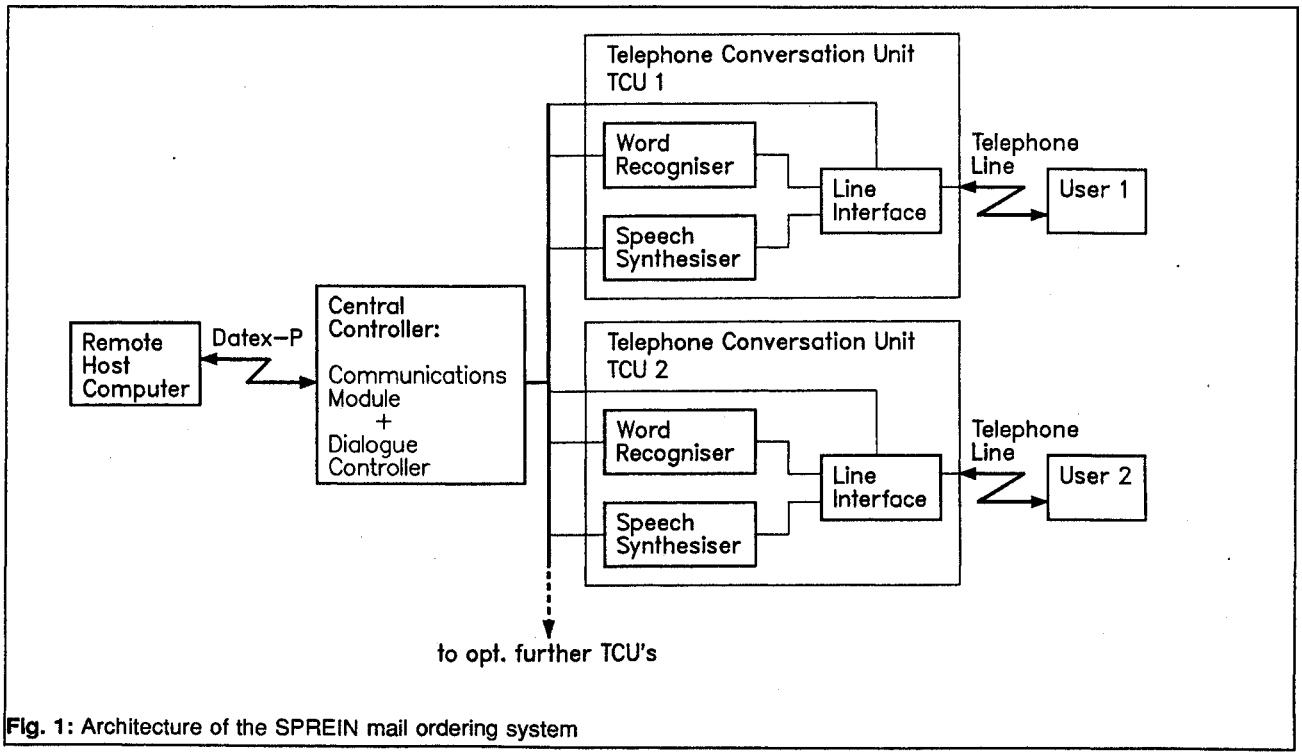

Fig. 1: Architecture of the SPREIN mail ordering system

10.21437/Eurospeech.1991-226

Two competitive systems were developed, one of them the Phlips system described here. The systems were evaluated in a one year's field trial in the Hamburg area in cooperation with the mail order house OTTO-Versand.

2. System Architecture and Hardware

The system structure of the SPREIN voice server is given in fig. 1. For each telephone line to be served, a telephone conversation unit (TCU) consisting of a speaker independent isolated-words recogniser, a speech synthesizer, and a line interface exist. A central controller interfaces to the remote computer of the mail order house, controls the TCU devices, and handles the dialogues for each TCU. In the trials, a configuration employing two TCUs was used.

The **central controller** is connected to the order computer via a data link running the widely used EHKP protocol /1,/ (among other services for BTX, similar to the French MINITEL service). For this reason, the order computer needs no special communication protocol for SPREIN, but sees it just as special BTX terminal. To manage several parallel orders, the BTX connection is multiplexed.

Being responsible for the user interfaces, the central controller runs the dialogues for all connected TCUs on a 68010 Microprocessor under UNIX. Although UNIX is not a real-time operating system, response times do not cause any perceptible delay. UNIX allows a comfortable and modular implementation of the software in form of separate processes for each major task, running independently and communicating only via pipes.

In the telephone conversation unit, the **line interface** detects incoming calls and other line signalling, holds a hybrid to separate received and transmitted speech signals, and a gain adjust to control volume of incoming speech signals.

For user guidance and feedback, speech is output generated by a **phoneme synthesizer** SAMT4. The synthesizer was developed and built by the research institute of Telekom /2,3/. Its intelligibility is worse compared to natural speech (82% to 93% intelligibility measured with the Sotschek monosyllabic rhyme test according to /4/), but the trial intended to test acceptance of synthetic speech, too.

Output messages for the synthesizer were prepared such that in a first step, the intended message was converted into phonemes and prosody information using automatic tools. After specification of all messages, they were further optimized for increased naturalness and intelligibility at the Telekom research institute. Although this way of output speech preparation takes some time, it avoids concatention problems, that typically occur for large sets of natural speech messages.

3. Speech Recognizer

The speaker independent isolated-word recognizer runs on a 68020 microprocessor system supported by a TMS 32010 DSP system for feature extraction. Recognition is done by a pattern matching approach. Feature vectors consist of 16 spectral components, and both references and recognition patterns are compressed to 16 vectors by trace segmentation. A gap of 0.4 sec after each input word is necessary for endpoint detection. Algorithms and software are described in more detail in /5/.

For the mail order application, only small vocabularies are necessary, consisting of either the 10 digits or 'yes'/'no' and some control words (stop, cancel, pause, repeat, correction), with at most 14 words active for recognition. A speech corpus containing 45 male and 55 female speakers uttering each word at least four times was collected for training.

The recognizer uses a male and a female reference template for each word, generated off-line. In one version, two reference templates were used for some digits per sex. Results for the different recognizer versions will be given later.

Our recognition results showed that at least for the used recognizer, about 100 training utterances per word collected from at least 25 different speakers are sufficient to create speaker independent references for one sex, as for training sets of this size, error rates for recognizing the training set and an unknown test set did not differ significantly any more.

4. Dialogue Concept

The dialogue design had to follow two contradicting guidelines. On one hand, it is wellknown that, due to imperfect technology and uniqueness of voice I/O, voice user interfaces have to be tailored as close as possible to the application. On the other hand, the system was demanded to be easily adaptable to other applications.

To resolve these contradictions, SPREIN was designed as an intelligent server. This means that the remote application controls the general states of the application, whereas SPREIN hs to be knowledgeable how to convert application demands into a userfriendly man-machine-communication.

For example, only the order computer knows whether an article number is related to an entiety countable in pieces or in meters, but the computer does not know whether the terminal will present demands to a user in a visual way or by voice. On the other hand, SPREIN does not know what kind of data the user will next be asked for, or whether there is an underlying syntax for order numbers that might be used to increase recognition accuracy, but it knows what to ask the user for each application demand, how to supply help, and how to cope with input and recognition errors.

To allow such a rather application independent dialogue concept, most application specific information is put into application specific data files, and furthermore, a protocol for communication of order computer and SPREIN system was specified, able to provide detailed format information by referencing to elements of the data files. As data files are in text form and read at startup of the system, it may simply be adapted to similar application by changing data files and restarting the system.

4.1 Communication SPREIN - Order Computer

The order computer talks to the SPREIN system via **item prompts**, and SPREIN answers by sending the requested data. An item prompt consists of an optional message field, a reference to a prompt unit, a response vocabulary, and an input format specification.

The optional **message field** refers to a set of messages directly related to the ordering procedure to supply information like "This article is not on stock." to the customers and may adress up to 512 predefined messages.

Prompt units refer to meaningful control structures like 'entry of an account number' or 'entry of item confirmation'. The control structures contain announcements for user guidance, recogniser control information and flow-of-control information.

A **response vocabulary index** specifies the vocabulary subset to be used in the input data field to be echoed back to the order computer. The active vocabulary for recognition is determined by activating the selected response vocabulary plus additional control words with only SPREIN-internal meaning like 'pause', 'repeat', 'cancel' etc.

The **input format specifier** determines minimal and maximal length of the input string. E.g., a customer's account number may vary from seven to nine digits, whereas a confirmation is always a one-word entry.

4.2 Communication SPREIN - User

For customers, every item prompt is structured in a similar, consistent way and generally consists of three subunits:

- a set of prompt messages for the first word entry, giving more general information about the data to be input, and allowing additional control words to e.g. pause or repeat an order computer's anouncement,
- a set of prompt messages for entry of trailing digits, including facilities to correct recognition errors,
- and an echo of a complete digit string including a confirmatory yes/no question.

Confirmative questions asked by the order computer are treated similar to a mixture of first and last subunit.

Dependent on whether the user identifies himself as being new or accustomed to the system, he will get either sentence-long, polite prompt messages or only some key-words for each new item prompt. If the user wants more help, he has to keep silent, and then goes though several levels of additional explanations and help instructions until finally either some input word is spoken or a transmission line breakdown is assumed. After a digit entry, the recognition result is echoed, and new input is expected. A short dialogue example shows the different help levels:

SPREIN: Please give your account number
USER: *<says nothing>*
SPREIN: Use "Cancel" or "Stop" to finish the order. If you want to have a break, please say "pause".
USER: *<says nothing>*
SPREIN: You can repeat announcements now using "repeat". Else utter the first digit of you account number now.
USER: *<says nothing>*
SPREIN: I cannot hear you. Could we have a transmission problem? Please speak again.
USER: 1
SPREIN: 1
USER: 2
SPREIN: 2 *<and so on>*

To correct errors, either the whole string may be deleted, or the user may choose to change only the last digit. For digit correction, SPREIN checks the wrong digit for a second choice and requests a confirmation if the second choice's distance is near to the distance of the optimal match. An example:

USER: zwei ! *this pronounciation of '2' is not allowed*
SPREIN: drei ! *because of its similarity to '3'*
USER: correction
SPREIN: Did you say 'zwo' then? ! ! *accepted* ! *pronounciation of '2'*
USER: yes
SPREIN: 2 ! *standard dialogue: echo of last digit*

We found that, due to typical confusions (in German e.g. 0 - 9), in about 90% of all cases the runner-up is the correct choice for a misrecognition. Especially for people causing systematic recognition errors due to dialect or for pathologic reasons this is a valuable aid, because in these cases a repetition of a word will not necessarily lead to better recognition.

5. Recognizer Evaluation

Several versions of the speaker independent recognizer were evaluated off-line in 1988 to 1990, both internally and by Telekom. The Telekom evaluation used 2600 words spoken by 100 male and 100 female speakers from all over Germany. Each word was presented three times:

- directly from a digitally recorded tape
- with a local telephone line switched between recorder and recognizer
- with a long distance telephone line switched between recorder and recognizer

The first recognizer version achieved 3.4% error rate for the digits in internal tests, but 6.9% in the Telekom evaluation in spring 1988. This version was not accepted as an error rate of 5% was demanded. As cause for the low recognition rate, it turned out that the special telephone line used for automatic training data collection had a low pass characteristic with an attenuation of 16 dB at 2 KHz.

After setting up a training database with a flatter frequency response, a new evaluation of the digits resulted in 1.9% error rate in internal tests and 2.7% average error rate in the Telekom evaluation (see Tab 1.), qualifying the recognizer for the field trial. In the field, the error rate rose and was estimated to be three times higher than the measured error rate.
A major part of the misrecognitions, both in the Telekom evaluation and in the field, was caused by digits being mistaken for '3', '5', or '9', happening to be those references with more than one reference template per word per sex. After increasing the training data base and restricting references to one template per word per sex, internal recognition rate dropped by 0.1%, but the average error rate of the Telekom evaluation in aug. 1990 went down from 2.6% to 1.8%, leading to the surprising result that the recognizer performed better in external than in internal tests. Maintaining the ratio of three between field and off-line test, errors rate in the field came down to about 6%.

	errors (%)	rejections (%)	total errors (%)
nov. 1988 evaluation for digits			
internal test	1.9	-/-	1.9
ext.: no line	2.0	0.1	2.0
ext.: local line	2.5	0.0	2.6
ext.: long dist. line	3.4	0.2	3.5
ext.: average	**2.6**	**0.1**	**2.7**
aug. 1990 evaluation for digits			
internal test	2.0	n.c.	2.0
ext.: no line	1.1	0.0	1.1
ext.: local line	1.8	0.0	1.8
ext.: long dist. line	2.5	0.0	2.5
ext.: average	**1.8**	**0.0**	**1.8**
aug. 1990 evaluation for control words			
ext.: no line	0.1	0.0	0.1
ext.: local line	0.1	0.0	0.1
ext.: long dist. line	0.1	0.0	0.1

Tab. 1: evaluation of recognizer error rate

6. Field Trial Results

In 1989 and 1990, SPREIN was evaluated with about 40 test persons, all of them aquainted to several forms of mail ordering (BTX terminal, mail, human telephone operator), too. The test persons were introduced to the trial and to the voice ordering system. A polling institute surveyed the field trial by interviewing the test persons at three different times. The results can be summarized as follows:

Speech synthesis:
The synthetic voice is accepted as clear and intellegible; about 80% judge the voice quality as 'good' at the end of the trial. 70% judge the talking speed as adequate; but in some cases it seems to be too slow.

Speech recognition:
At the end of the trial, only few users still have problems with some specific digits to be recognized. Averaged over all questionaires, 29% feel that recognition accuracy should be improved. For speaker specific error rates higher than 10%, the acceptance drops significantly. Misrecognitions of control words, which occurred very infrequently, was very annoying.

Dialogue concept:
The dialogue concept confuses new users, but is accepted and judged to be sensible after several orders. Most of the help information is needed only by new users. Offering a second choice for misrecognition took some time to get used to, but was very helpful to people with systematic misrecognitions.

General performance:
- An average SPREIN order lasts more than three times as long as human operator ordering. So 41% of the test persons complained about operating speed.
- Acceptance of the system was improved significantly by the fact that the system was on-line 24 hours, whereas human operators were not available at night and at weekends.
- During the one-year field trial, 40% of the test persons kept on using SPREIN (due to technical problems affecting availability of SPREIN, about half of the users did not participate actively in the trial any more after a year. But the remaining users demanded SPREIN to stay on-line after the end of the field trial.
- A foreigner with a rather severe accent noticed that the system had less problems than an operator to understand him, and that SPREIN was more patient.

7. Conclusions

Most surprising for speech researchers is, that not the performance of speech synthesis or speech recognition decides about acceptance of an application, but whether the users benefit from the application. For applications, this means that it is not sufficient to replace just some existing I/O media by voice, but either the application should be new, or an existing application should at least be significantly improved, e.g. by adding new features.

Applications using isolated word recognition should not target to collect large amounts of data by voice, as has been done in the SPREIN application. To compete with a human operator in speed of data entry, technologies like connected-word recognition, echo cancellation for permantent interruptability of announcements, and a dialogue concept giving more freedom to a user have to be adopted.

References:

1. Deutsche Bundespost: Bildschirmtext-Rechnerverbund Protokoll-Handbuch. Deutsche Bundespost, 1985
2. H.E. Wolf: Entwurf und Realisierung eines Formantsythetisators mit paralleler Filterstruktur für die Sprachsynthese nach Regeln. Dissertation Th Darmstadt, 1981
3. H.E. Wolf: Control of Prosodic Parameters for a Formant Synthesizer Based on Diphone Concatenation. Proc. ICASSP 81, pp. 106-109, 1981
4. J. Sotschek: Sprachgüteuntersuchungen an einem Sprachsynthetisator mittels Reimtest-Verständlichkeitsmessungen. Proc. 6. FASE - Sopron, Hungary, 1986
5. M.H. Kuhn, H.H. Tomaschewski: Improvements in Isolated Word Recognition. IEEE Trans. ASSP-31 No.1, pp. 157-167, 1983