2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94) Kyoto Research Park, Kyoto, Japan September 26 - 27, 1994

EXPERIENCE WITH THE PHILIPS AUTOMATIC TRAIN TIMETABLE INFORMATION SYSTEM

Harald Aust, Martin Oerder, Frank Seide, Volker Steinbiss

Philips GmbH Forschungslaboratorien
P.O. Box 1980, D-52021 Aachen, Germany

ABSTRACT

We introduce an automatic system for train timetable information over the telephone that provides accurate connections between 1200 German cities. The caller can talk to it in unrestricted, natural, and fluent speech, very much like he or she would communicate with a human operator, and is not given any instructions in advance.

In an ongoing field trial, this system has been made available to the general public, both to gather speech data and to evaluate its performance. This field test was organized as a bootstrapping process: initially, the system was trained with just the developers' voices, then the telephone number was passed around within the department, the company, and finally, the outside world. After each step, the newly collected material was used for retraining and general improvements.

The observations and results from this test are reported here.

1. INTRODUCTION

We have developed a prototype of a system that people can call in order to obtain information on the schedule of the German railway system.

While inquiry systems for similar tasks are already commercially available, they are not very user-friendly. Callers have to interact with them either by pushing keys on their touch-tone telephone, or by uttering one of just a few words the system can understand. The ensuing dialogue is usually menu-driven, rigidly structured, and accompanied by lengthy explanations.

By contrast, our system allows users to talk in unrestricted, natural, and fluent speech, very much like they would converse with a human operator. They do not receive any instructions beforehand.

The basic ideas of this system, which is described in more detail elsewhere [2, 3, 8], are as follows: Speech recognition, speech understanding, dialogue control, and speech output are separated into distinct modules that are executed sequentially. Well-defined interfaces ensure that the system can be easily maintained and that entire modules can be exchanged without affecting the remaining parts.

The speech recognizer creates a word graph and passes it on to the speech-understanding component. It is then parsed with a stochastic attributed context-free grammar, which is used as a language model, to identify the relevant parts, and to compute their meaning; this parse does not necessarily cover the entire sentence, so grammatically incorrect spontaneous utterances, as well as recognition errors, can be handled properly. A dialogue manager analyses the results and either comes up with a new question or accesses the timetable database. The answer to this query, as well as the questions, are converted to spoken language by concatenating and replaying appropriate pre-recorded phrases.

This prototype is a pure software solution (written in C) that runs on our Unix workstations. As a consequence, response times can vary with the current load on the computers, but we usually attain real-time operation. The system is connected to the public telephone network, which means that it can be called from all over the world. It is, however, still a research prototype that is in no way interrelated with the official railway information service of the Deutsche Bahn AG. The database we use, though, is a real one, so the information a caller gets is accurate.

The system understands and responds in German; all examples in this article were translated into English. The vocabulary of the speech recognizer currently consists of approximately 1800 words, 1200 of which are railway station names. They are made up of the most important German cities, plus a number of smaller stations in the Aachen area and a few from abroad. Since there are almost 10000 railway stations in Germany, currently not all connections are available. The above described selection, however, ensures that more than

This work was in part supported by the German Ministry of Research and Technology (BMFT) under contract No. 01 IV 102 B.

95% of the timetable inquiries occuring in real life are covered.

A typical call to the system goes as follows:

System: *Good morning. This is the automatic timetable information service. How can I help you?*
Caller: *Good morning. I need a connection from Hamburg to Munich.*
System: *When would you like to go from Hamburg to Munich?*
Caller: *Next Friday, er, at eight o'clock.*
System: *So you want to travel on Sunday the 25th of September at 8 am?*
Caller: *No, on Friday.*
System: *So you want to travel on Friday the 23rd of September at 8 am?*
Caller: *Exactly.*
System: *There is the following connection: With IntercityExpress 793, departure from Hamburg at 8:02, arrival in Munich at 14:05. Would you like me to repeat the connection?*
Caller: *No, but I need a train for the return trip on Sunday.*
System: *At what time would you like to go from Munich to Hamburg on Sunday the 25th of September?*
Caller: *I have to arrive in Hamburg late in the evening.*
System: *So you want to arrive in Hamburg between 8 pm and 11 pm?*
Caller: *Yes.*
System: *There are three connections. This is the first one: With IntercityExpress 880, departure from Munich at 15:54, arrival in Hamburg at 21:58. Would you like me to repeat the connection?*
Caller: *No, thank you.*
System: *Would you like the next connection?*
Caller: *No, thanks, this one is fine.*
System: *Do you need another connection?*
Caller: *No.*
System: *I wish you a pleasant journey. Thank you for calling. Good bye.*

Just as in the communication process between humans, recognition errors and misunderstandings may occur. The system therefore allows the caller to correct these mistakes, as shown in the example. Under adverse circumstances, it may still happen that a customer cannot reasonably be understood. This will eventually be detected, and he or she will be referred to a human operator.

2. FIELD TEST

For several reasons, it was important for us to make many people call our system. First of all, we had to acquire training material for the recognition and understanding parts and to adapt the vocabulary accordingly. Secondly, we needed realistic calls that would allow us to debug the system, as well as to evaluate the dialogue flow and the system's overall performance. Furthermore, we wanted to learn about customers' demands and ideas, their general behavior and problems, so we could increase the system's usability and usefulness. And finally, we were interested to see in how far such an automatic system would be accepted.

For finding answers to questions like these, Wizard-of-Oz (WOZ) scenarios in which a human being takes the part of the computer are widely used [4, 7]. While such a set-up is well suited for the collection of speech data, the other aspects cannot be covered as well. Obviously, the system cannot be tested, debugged and evaluated if replaced by humans during the test. Also, the way people react to a system and their attitude towards it largely depends on its performance. This in turn is closely related to the recognition and understanding rate which is hard to simulate. Therefore, in a WOZ environment, the utterances collected, as well as the observations made and comments given, are less realistic than when the actual system is used.

Because of these reasons, we have conducted a field test with our system. The general idea has been to tell people about the existence of a new automatic inquiry system and encourage them to use it or at least to try it out. They were not given details or instructions in any case. It should be noted, though, that while our approach allows for a rather exact evaluation, the collected data is not necessarily completely realistic: people who call in because they are curious tend to behave differently from those who really need an information.

The field test was organized as a bootstrapping process. We began with a speaker-dependent system that was not to be used over the telephone but with an ordinary high-quality microphone, and that displayed its output on the computer screen. Our initial training material for both speech recognition and speech understanding consisted of roughly 1000 sentences that a number of different people had thought up. This data, however, did not account for diverse dialogue situations: about 85% were — not necessarily complete — single-sentence inquiries for train timetable information ("on Monday at eight I want to go to Stuttgart"), while the remaining 15% were only slightly related to the conversation subject ("when will the schedule change"). As a consequence, several words and phrases that are often used in real dialogues were poorly rec-

ognized. The best example is the indispensable word "yes" which in the beginning was hardly understood at all since it did not occur in the training material.

With this original system, some speech data was collected that had mostly been spoken by the developers themselves. Having trained the system anew, a telephone version was installed that showed, foreseeably, poor performance. In order to induce other persons to call, the telephone number was first circulated within our research group, and later inside the department and the entire laboratory. After each step, the newly compiled material was harnessed for retraining and general improvements.

Motivating other people to actually call our system turned out to be a major problem. We used personal talk and posters for advertising. Unlike described in [7], we always pointed out that this was an automatic system, hoping to raise curiosity in this way. To avoid that only the relatively small number of colleagues actually planning a journey would call, we sought to explicitly encourage everybody to simply try the system and see how it worked.

While we received clearly less calls than we had expected, those that came in eventually enabled us to improve the system's performance to a point when we were ready to address the general public. Beginning in February 1994, it has been promoted through press releases and radio interviews. The number of incoming calls showed considerable peaks after each publication; often, the system was not idle for more than one or two seconds before the next call arrived. Since we were offering only a single telephone line, we probably missed many calls, in particular because people who call the system only out of curiosity are likely to lose interest if they attempted several times to get through but always heard the busy signal.

Proceeding in the above described way has permitted us to debug, evaluate, and improve our system ever further. The disadvantage is that many of the incoming calls were made by people who only tested the system and did not really need an information. We would probably have to connect our system to an official information service to overcome this drawback.

3. CALL TRANSCRIPTION

All incoming calls are recorded and transcribed manually. This is a considerable expenditure of time, but a transcription of the speech data is necessary for further training of both the recognition and understanding components. Besides, the dialogue flow can be evaluated, and possible problems or awkward passages can be detected at the same time.

Each call is annotated with attributes that describe its particular properties; voice characteristics (male/female/child), speaker accent, and line quality (low volume/noisy/background talk/background music etc.) being among the most prominent. We also note whether the call was successful or not, which cannot always be rated because some callers hang up immediately or check out the system instead of seeking a train connection.

The system itself always stores the output it creates, i.e. the questions and the query results, so together with the transcription, the full course of a dialogue can be recovered. It is helpful for the transcriber to see what kind of system action lead to an otherwise unexplicable user response, and in a future version, this information may also be useful for dialogue-situation specific training.

The recognizer's vocabulary is updated automatically. It consists of the words the speech understanding component can process, plus all of those that occured in callers' utterances at least n times, with n typically being between 2 and 4, depending on the desired vocabulary size.

4. EVALUATION

While it is trivial to evaluate a stand-alone speech recognizer, and feasible to do so for a system that creates a database query from a single sentence, there is no easy way to evaluate dialogue systems automatically [1]. After all, the all-important criterion is user satisfaction, which cannot be determined as easily as, say, the word error rate.

Probably the best approximation is to measure the percentage of callers who obtained the information they asked for. This, of course, is only a simplified approach since important aspects like the speed of the system, clearliness and understandability of the voice output etc. are not accounted for.

To get at least a rough idea of callers' opinions, we ask them for general comments and for suggestions for improvements at the end of a successful dialogue. Their remarks are recorded but not processed by the speech recognizer. Detailed questionaires as used in [5] or [6] could certainly yield more valuable information, but our experience shows that in an anonymous telephone environment most callers would not bother to answer them. Indeed, only very few people even respond to our short invitation to comment on the system. Those who do, however, are usually enthusiastic; phrases like "very good" or "wonderful" abound. Negative statements mostly refer to difficulties that occured during the previous dialogue and are hardly ever of a more general kind.

It is interesting to see that evidently many people are not aware that speech recognition is a technology that does not yet work 100% accurately. Problems caused by recognition errors are often not perceived for what they are but are attributed to general weaknesses of the system. This impression is confirmed by a number of comments explicitly referring to the speech output: it seems that in the minds of the general public speech output, and not speech recognition, is regarded as the real challenge.

Another remarkable observation from the feedback is that the satisfaction with and the approval for the system is apparently less dependent on the course of the previous dialogue but rather on the general attitude of the caller. This may, however, be specific for Germany where people tend to be critical towards new — and impersonal — technologies.

These experiences indicate that a good and human-like speech output is important for a high acceptance of an automatic inquiry system. Hence our method of using a human voice instead of a synthesized one seems to be a good choice.

5. OBSERVATIONS AND PROBLEMS

In this section, we are going to report on the observations and problems most frequently noticed during the field test.

5.1. Technical Problems

A sizable number of difficulties was caused indirectly by the telephone interface. Our current installation is very simple and does not allow a flexible adaptation to an individual person. We consider a caller's utterance to be finished after a predefined amount of silence went by. If this pause is chosen to long, an awkward delay will result. On the other hand, if it is too short, people speaking hesitatingly may be interrupted in the middle of a sentence.

We cannot compensate for different volumes of callers' voices, either. While some of them spoke so quietly that the system could not find out whether they talked at all, others were so loud that considerable distortions were caused already by their receiver's microphone. Unfortunately, there is little that can be done about that.

The detection and processing of the actual speech signal was occasionally made more difficult by background noise, e.g. from a radio or television set. Some callers even created their own distractions by talking to other people in the room: "Amazing — he understood everything!" (this astonished remark then confused the system completely, and the call was finally aborted).

5.2. Problems with the Database

Two problems that we had not anticipated were caused by the database. Firstly, because we create speech output by replaying pre-recorded phrases, everything that is to be converted from written to spoken language must be known in advance. This, of course, is especially true for all timetable information possibly found in the database. We had not expected to encounter words like "ferry" or "footpath" when asking for train connections and consequently could not output them.

Secondly, whenever we received a new release of the database because of the seasonally changing train schedule, we found that some stations did not exist anymore or were named differently.

The solution here would be a closer cooperation with the manufacturer of the database.

5.3. Recognition Errors

While much of the natural language understanding and dialogue research efforts in the past went into systems that accept typed input, our system by definition deals with spoken language. The major difference is that in our case recognition errors can and do occur. In fact, because of the adverse conditions of speaker independence, spontaneous speech, low signal quality, open and relatively large vocabulary, and real-time constraints, the word error rate of the recognizer is currently about 25%.

In an inquiry system, recognition errors can have two effects: something said by the caller may not be correctly understood, and something he or she did not talk about might be erroneously found in the speech signal. The first case is not too troubling if it doesn't occur too often — an appropriate question will cause the caller to repeat what he or she said —, whereas the second turns out to be more severe. People tend to be confused if the system "knows" something they have not yet talked about, and may not be able to correct such a misunderstanding. Therefore, the dialogue should be designed in a way that at any particular time it can only understand those things that a caller can be reasonably expected to say.

The disadvantage of this strategy is that there will always be people who say something that makes sense in certain circumstances, but the system does not comprehend it. On the other hand, a caller cannot distinguish such a situation from a simple recognition or understanding problem, and these are unavoidable, anyway.

A typical example where we changed the dialogue in order to account for the recognition difficulties outlined above is the way we handle verifications. Due to

the high error potential, it is vital for an automatic inquiry system to verify what it believes it understood by asking appropriate questions. Otherwise, an incorrect query may result.

Because we did not want to disrupt the normal flow of conversation, our first approach was to come up with a single verification question like "So you want to go from Hamburg to Munich tomorrow at 3 pm?" after all information was gathered, and to allow the correction of all data at this point. Unfortunately, the consequence of this strategy often was that a confirmation given by the caller, like "Yes, exactly", was misrecognized as correction, e.g. "at 4 pm". The result was in some cases that while the system had originally understood everything perfectly, the unintended correction caused severe problems that finally lead to a wrong query or an aborted call.

In our current system, we therefore verify everything by changing the subsequent question appropriately, as in "When would you like to go to Munich?" instead of "When would you like to go?" when the destination "Munich" is to be confirmed. A correction is only possible for those values that appear in the question (in the example, the destination) which greatly reduces the odds of misunderstandings. This strategy has a disadvantage, too: some people do not realize that they can correct a value since they are not explicitly asked to do so.

Another major difference between written and spontaneously spoken input lies in the correctness of the utterances. While many of the observed expressions are grammatically correct sentences or at least parts thereof, word sequences like "twelve midnight for to be in Hamburg" also appear. This underlines the validity of our understanding approach that only looks for meaningful words and phrases, regardless of their order within an utterance.

5.4. Reactions to Questions

We found that the way a particular question is asked greatly affects the response of the caller. As shown in the example in Section 1, our system originally initiated the dialogue with "How can I help you?". This formulation, natural as it may be among humans in a similar situation, apparently confused several people who did not know how to react. When this phrase was altered to a more suggestive "From where to where do you want to go?", quite often, simple answers like "from Hamburg to Munich" were the result.

Of course, there is no guarantee that questions will always be answered in the way one would expect. In fact, we often encountered situations in which a response did not only not answer the question but was, at least at first sight, altogether illogical:

System: *When would you like to go to Hamburg?*
Caller: *No.*

We suppose that in this and comparable situations the callers tried to speak in a machine-like language — in this case, to indicate that they did not want to go to Hamburg — because they did not realize that the system would have understood a response like "No, not to Hamburg, I want to go to Munich."

5.5. Callers Who Only Test the System

Several callers apparently did not need an information, and did not even pretend they would, but tested the capabilities and limitations of our system. Typical of this is the use of absurd phrases like "I want to go yesterday", "on the 30th of February", or "from Hamburg to Hamburg". Also, some people were ingeniously coming up with alternatives for the word "yes": "right", "okay", "perfect" are a few of them.

We do not think that these are problems that should be spent major effort on. After all, a realistic caller wants an information and can therefore be expected to be cooperative. Besides, it will never be possible to make a system absolutely foolproof anyway.

Another phenomenon was that an unexpectedly high number of stations was asked for that were not in the vocabulary. The explanation is that people who only wanted to try the system often asked for a connection to their home town, or purposely tried small stations to see whether they would be understood. As mentioned above, the selection of the stations in our vocabulary should ensure that more than 95% of real inquiries can be answered.

5.6. Accents and Dialects

We were amazed to observe that even people speaking in a German dialect or with strong foreign accents almost never had problems to be understood. This may partly be due to a particularily careful and emphasized pronunciation typical of non-native speakers.

Occasionally, though, foreigners used phrases that native speakers would never consider and that were not covered by the system's grammar. Consequently, they were not understood, even though it would have been clear to humans what they meant. However, this does not constitute a real problem: if an automatic inquiry system were to be employed in an environment where many non-native speakers could be expected to use it, the appropriate phrases could simply be added to the grammar.

5.7. Diverse Opinions on Details

Another point that we had not expected is how widely the opinions of people we asked about our system varied, even on very specific topics. To give a few examples:

- Before our system became fast enough for real-time operation, a piece of (electronic) music was played whenever the caller had to wait for a response. About half of the people asked liked this music very much, the other half detested it.

  We found, though, that some kind of pause-bridging sound is necessary if a system cannot answer immediately. Otherwise, callers are invariably confused and wonder whether the line was interrupted.

- While some people found the system's announcements too slow, others complained about their high speed.

- Comments on the (male) voice of our system range from "very pleasant" to "boring" and "should be changed".

The consequence of this observation is that one has to be very careful when it comes to modifications of the system. In particular, it should not be changed because of individual opinions, not even if they are identical to one's own.

6. RESULTS

Currently, we receive about 1000 calls per month; altogether, we have collected more than 5000 so far. Approximately one third of them cannot be used for evaluation purposes since they were made by people who only played with the system, used it for party entertainment, or hung up right after the initial greeting. Of the other two thirds, 10% seem to consist of real requests, while 40% of the callers apparently only try the system. For the remaining 50%, this cannot be decided. The success rate for these three groups averages about 75%. One quarter of the remaining calls fails due to poor recognition performance, which we hope to improve further as we collect more training data. The rest is asking for stations that are not in the vocabulary, or has other problems with the dialogue. We are confident that we can achieve a success rate of 90% within a year.

7. CONCLUSION

We have described our experiences with an automatic train timetable information system that was made publicly available in a broad-based field trial. This test was organized as a bootstrapping process, which allowed us to start with only little original training data and utilize the incoming calls for improvements from the very beginning. We found that this is a good way to collect training material and to evaluate the system at the same time.

The success rates achieved in the test make clear that the underlying technology is well-suited for realistic applications. Most importantly, the reactions of many callers show that automatic systems of this kind are accepted and welcome since they can often provide better and more easily accessible service.

8. REFERENCES

[1] J.F.Allen: *Session 1: Evaluating Spoken Language Systems.* In *Proc. Speech and Natural Language Workshop* pp. 5–6, DARPA, Morgan Kaufmann Publishers, San Mateo, CA, February 1992.

[2] H.Aust, M.Oerder: *Generierung einer Datenbankanfrage aus einem gesprochenen Satz mit einer stochastischen attributierten Grammatik.* To appear in *Proc. MUSTERERKENNUNG 94*, Springer-Verlag, 1994.

[3] H.Aust, M.Oerder: *Database Query Generation from Spoken Sentences.* In these proceedings.

[4] B.Bly et al.: *Designing the Human Machine Interface in the ATIS Domain.* In *Proc. Speech and Natural Language Workshop* pp. 136–140, DARPA, Morgan Kaufmann Publishers, San Mateo, CA, June 1990.

[5] E.Gerbino et al.: *Test and Evaluation of a Spoken Language Dialogue System.* In *Proc. ICASSP 93* pp. II-135–II-138, Minneapolis, MN, 1993.

[6] C.T.Hemphill, J.J.Godfrey, G.R.Doddington: *The ATIS Spoken Language Systems Pilot Corpus.* In *Proc. Speech and Natural Language Workshop* pp. 96–101, DARPA, Morgan Kaufmann Publishers, San Mateo, CA, June 1990.

[7] R.Moore, A.Morris: *Experience Collecting Genuine Spoken Enquiries Using WOZ Techniques.* In *Proc. Speech and Natural Language Workshop* pp. 61–63, DARPA, Morgan Kaufmann Publishers, San Mateo, CA, February 1992.

[8] M.Oerder, H.Aust: *A Realtime Prototype of an Automatic Inquiry System.* To appear in *Proc. ICSLP 94*, Yokohama, 1994.