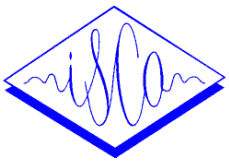

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

EXPERIENCES FROM A REAL-WORLD TELEPHONE APPLICATION: teleDialogue

Per Rosenbeck & Bo Baungaard

Jydsk Telefon, Research & Development Department
Sletvej 30, 8310 Aarhus-Tranbjerg, Denmark

ABSTRACT

Unpredictable challenges emerge from the task of moving speech technology applied in laboratories to applications working in the telephone network.

This paper summarizes experiences and results achieved from a real-world telephone application that makes use of an advanced user-dialogue and recognition of isolated words using the Continuous Hidden Markov Model approach.

The service developed at Jydsk Telefon, Denmark, is named teleDialogue. It has been running as a commercial service for a field trial period of one year (May 91 to May 92), serving more than 1 million subscribers. TeleDialogue enables call establishment to six commonly used telecommunication services in Denmark simply by pronouncing the names of these services. In this way teleDialogue integrates these six services into one that can be reached by calling one specific service number (0020).

In order to achieve information on the use of teleDialogue, it is designed to give comprehensive statistics on the use of the service. Furthermore it is possible to monitor and even record user dialogues with the service.

Even if the recognition performance in the laboratory is higher than 95% (SAM standard), the successful transaction rate in the real-world drops to 85%, meaning that 85 of 100 initiated calls are directed to the desired service.

A sample material, obtained by listening to user dialogues, has been selected to reveal the reasons for this performance drop i.e. extraneous speech, background noise, utterance input level and utterances given by children. Each of these subjects are discussed thoroughly in the paper.

Jydsk Telefon is, with teleDialogue, among the first in Europe to integrate speech technology into real-world telephone applications.

1 INTRODUCTION

The telephone application service named teleDialogue, described in this paper and in [1] and [2], is developed at Jydsk Telefon, Denmark. It has been running as a commercial service for a field trial period of one year (May 91 to May 92), serving more than 1 million subscribers all over Jutland. Using speech recognition, teleDialogue enables ordinary subscribers to do call establishment, simply by pronouncing the names of the following telecommunication services:

1. Klokken	(Time Service)
2. Horoskopet	(Horoscope Service)
3. Sportsavisen	(Sport News)
4. Vejrmeldingen	(Weather Forecast)
5. Telefonavisen	(General News)
6. Folketingsdebat	(Political News)

In this way teleDialogue integrates these six services into one service, that can be reached by calling one specific service number (0020). More than 110 000 subscribers have used teleDialogue in the field trial period.

When teleDialogue receives an incoming call it plays back a message: 'Welcome to teleDialogue', and waits for the user to pronounce, in an isolated manner, the name of one of the services. teleDialogue recognizes the name and confirms the choice by playing back the name of the service, while connection to the desired service is being established.

The purpose of teleDialogue was to collect experiences from a real-world speech technology application. These experiences are described in this paper. Chapter two gives a more thoroughly description of teleDialogue. Experiences and results are presented in chapter three. Summary and conclusion is given in chapter four.

2 DESCRIPTION OF teleDialogue

A number of requirements was set for the application to work in the telephone network:

1. Multi-user system.
2. Real-time performance.
3. Reliability.
4. Speaker-independency.
5. High recognition performance, regardless of age, sex and dialect.
6. Incorporation of a rejection technique for extraneous speech.
7. User-friendly man-machine interaction.

To give multiple users access simultaneously and achieve a real-time performance, the application is based on two Intel 386 PC's using the multitasking operating system iRMX. The use of two independent PC's gives the application redundancy and thereby

10.21437/ICSLP.1992-394

reliability. The PC's are connected to a public exchange via PCM lines (2 Mbit/sec.). Each PC is equipped with four digital signal processor (DSP) boards which perform the speech recognition. These boards are commercially available. A fifth board, which is developed for this particular application, includes a DSP for DTMF decoding, a digital crosspoint switch, PCM interface, and a microprocessor for communication with the public exchange. The DSP's are AT&T's DSP32C which perform 25 MFLOPS each.

Each PC is capable of handling 15 users, that is 15 incoming and 15 outgoing lines, and perform four recognitions simultaneously. The switching of the incoming and outgoing lines within the PC allows for access to a comprehensive statistical material on the use of the application. Figure 1 gives an overview of the teleDialogue system.

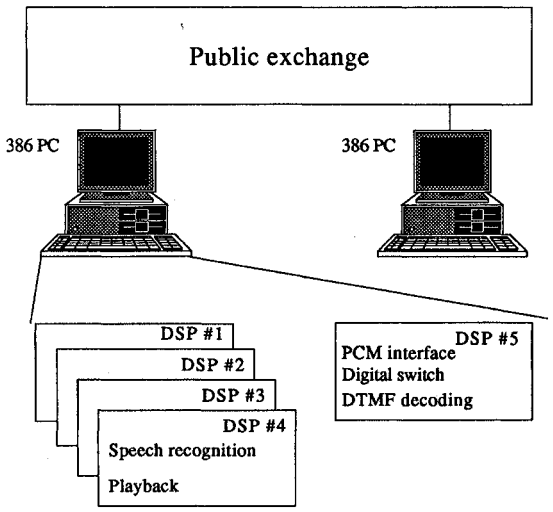

Figure 1 Overview of the teleDialogue system.

The fourth, fifth, and sixth requirement are met by using Continuous Hidden Markov Models (CHMM) for recognition of isolated words, where the words are the names of the services. Each Markov model consists of 10 states with one density function per state. The feature vector used for scoring the CHMM's consists of 8 LPC-derived cepstral coefficients and the 8 corresponding delta-cepstrum coefficients. Each model is trained with 64 speakers equally distributed regarding to dialects, age (above 16) and sex. The database for generating these models was recorded over dialled-up telephone lines. Although only six services are available in teleDialogue, 23 Markov models have been trained in order to cover a wide variety of synonyms, thereby increasing the robustness of teleDialogue.

In a real-world application, that uses isolated word recognition, it is necessary to be able to reject extraneous speech, i.e. non-vocabulary words and sentences. In teleDialogue this is done with garbage models [3], representing non-vocabulary words and sentences and a combination of these. The training data for the garbage models were collected during a small field trial.

The requirement of a user-friendly man-machine interaction is met by a well-designed dialogue. The dialogue was designed through an extensive use of the Wizard of Oz technique. This was done in an early stage in the development of the application. The dialogue is designed to handle skilled as well as non-skilled users. The latter group is informed on the use of teleDialogue by playback of a user guide and a user menu. The users are also encouraged to repeat, if they are not speaking in an isolated manner or if non-vocabulary words are uttered.

The dialogue begins with a short welcome message followed by a pause (2.5 sec.) where the skilled user can say a word and be recognized. If nothing is said in this recognition-window the user guide and user menu is played back to the user. Due to no talk-through facility, the user has to listen to the complete message (approx. 40 sec. of a nice and friendly female voice) or interrupting the message by using a DTMF tone. A 5 sec. recognition window follows, and if nothing is recognized, the user is prompted for a name. This is repeated three times before the call is disconnected. The transaction time defined as the time from observing a call to the time when the connection to the desired service is established, can be as low as 5 sec. for a skilled user. An overview of the user-dialogue is given in figure 2.

* "Welcome to teleDialogue"
* Recognition window 2.5 sec.
* User guide
* User menu
* "Enter the name of the desired service"
* Recognition window 5 sec.

Figure 2 Overview of the user-dialogue.

As soon as the user is recognized in one of the two recognition windows, a connection to the desired service is established. teleDialogue is free to serve other users, as long as the user is listening to the service. After six minutes teleDialogue plays back a warning to the user, telling that the connection will be closed in one minute. This is to avoid misuse or blocking of teleDialogue for other users.

3 EXPERIENCES AND RESULTS.

During the one year field trial period for teleDialogue, various types of experiences have been achieved on how to make an application reliable and robust. These experiences are divided into SAM assessment on recognition, real-life assessment and dialogue assessment.

I SAM Assessment on recognition.

Assessment of teleDialogue with 1457 utterances was performed according to the guidelines set up in the SAM ESPRIT project [4] and [5], to give the overall performance for the recognizer, shown in table 1.

Correct Recognized	Miss	Substitutions
95.1 %	3.6 %	1.3 %

Table 1 Assessment results on teleDialogue extracted from the SAM scoring documentation file.

Substitutions is regarded as being most critical, since these utterances results in a service different from the desired. Miss situations can be dealt with through the dialogue by encouraging the user to repeat or speak louder/lower.

The 23 Markov models, representing utterances of the names and the most common used synonyms for the six different services, is one reason for achieving this relatively high performance. Another reason is the collection of the database for the models through dialled-up lines, thereby including different line characteristics, telephone sets and background noise sources, that will also be met using teleDialogue. The distribution of the database regarding dialect, age and sex is also considered important. The following results reveals the lack of children in the database, since this group of users have substantially lower recognition performance than adults.

The three garbage models representing sentences and words outside of the vocabulary are responsible for lowering the number of substitutions, though, giving more miss situations. A specialized garbage model representing one single word very commonly used is also included, and verified to lower the number of substitutions.

II Assessment on Real-life Situations.

An assessment of teleDialogue with 1148 real-life situations shows an overall successful transaction rate of more than 85%, meaning that more than 85 of 100 initiated calls is directed to the desired telecommunication service. The reasons for this 10% drop in performance is explored in this chapter.

Three attempts to recognize the user is given in teleDialogue to reach the desired service. The distribution is shown figure 3.

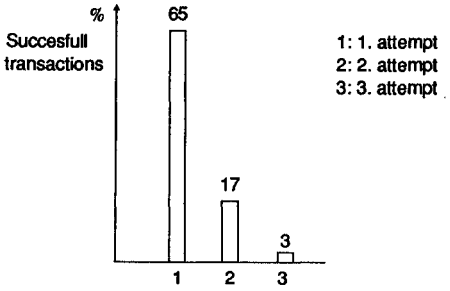

Figure 3 Distribution of successful transactions for first, second and third recognition attempt.

From the figure it can be seen that first attempt gives more than 76% of the successful transactions.

The main reasons for not being recognized in the first or second attempt is distributed as shown in figure 4.

Experiments have been conducted to set high/low values for energy in the utterances in order to lower background noise influence without affecting the recognition rate. If a certain utterance falls outside these values the user is encouraged to repeat louder/lower, thereby minimizing the risk of substitution errors. Another approach that have been tested and verified in a separate experiment is to include the changes in energy (delta-energy) in the feature vector, thereby making the recognition more robust.

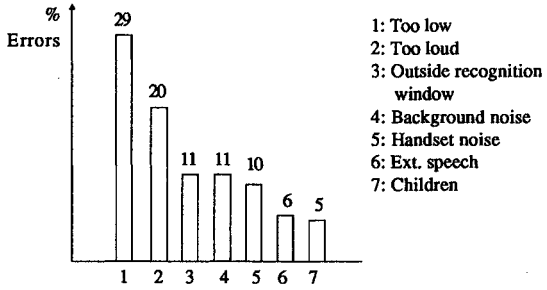

Figure 4 Distribution of reasons for the user not being recognized in first or second attempt.

The handset and background noise figures is also significant, even though noise effects to a certain degree is included in the Markov models, since the database was recorded through dialled-up lines. Background noise and handset noise could be modelled separately to lower these figures, but this approach has not been tested. Short noise pulses is taken care of simple by letting the endpoint detection ignore pulse duration less than 270 ms.

The number of misrecognitions outside the recognition window, in figure 4, is due to the user uttering while a message is played back. One way of lowering this number of misrecognitions could be to activate the recognizer 1-2 sec. before the end of the previous message. Obviously, this calls for a talk-through facility.

For more than 92% of the utterances that were not recognized in the first or second attempts, there have been a good reason for making the user repeat, thereby avoiding substitution errors.

The non-successful transactions are distributed as 7% of given-ups, defined as users interrupting the call instead of repeating the utterance. 5% substitution errors, and 2% false alarms. Finally 1% were not recognized at all. This distribution is shown for the 1148 real-life utterances in figure 5.

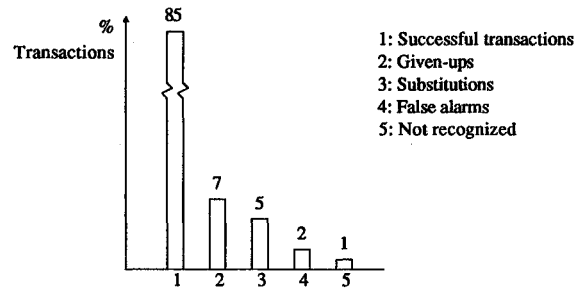

Figure 5 Distribution of 1148 real-life utterances.

The number of given-ups is highly dependant on a well-designed user-dialogue and a high recognition/rejection performance for the recognizer. Given-ups is 4.8% for adults and 2.6% for children. The number of substitutions is divided into 3.2% for adults and 1.3% for children, which together with given-ups indicates the lack of children in the database for teleDialogue. The low number of false alarms is due to the rejection capability of the garbage models.

III Assessment on User-interface.

The user-friendliness was tested by having 10 non skilled users try teleDialogue with the following success criterias. Results are presented on a scale from 1 to 5, where 5 is the best

* How easy to learn : 4.9
* How easy to use : 4.9
* How comfortably to use : 3.9

The conclusion was that teleDialogue is regarded as being very easy and comfortably to use.

As mentioned in chapter two, the dialogue is designed to handle skilled as well as non-skilled users. The user is encouraged to utter the name of the desired service in the 2.5 sec. recognition window, just after the welcome message. A questionnaire on the use of teleDialogue revealed that 81.5% of the users were using this short cut to the desired service after a few attempts. The transaction time can be as low as 5 sec. for a skilled user.

The announcement of teleDialogue in the press was done May 91, and again november 91. Not surprisingly these months gives peak values on the use of teleDialogue, see figure 6. Note also that there is a steady group of approximately 5000 users for a service like teleDialogue.

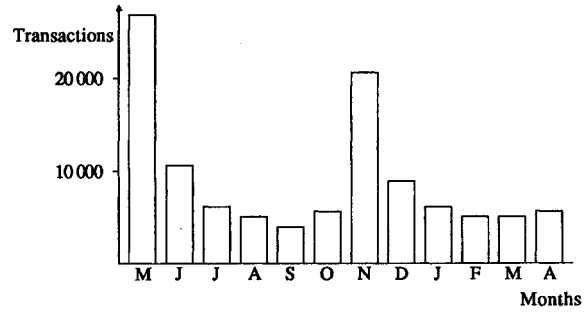

Figure 6 The use of teleDialogue in the field trial period from May 91 to May 92.

4 SUMMARY AND CONCLUSION

teleDialogue has proved to be a highly useful telephone application. The field trial has shown that people feels comfortably using this application.

With teleDialogue it is shown that a careful design of the user interface, is a crucial factor for the overall success of a telephone application. The Wizard of Oz technique made it possible to design an almost optimal user-dialogue at an early stage in the development of this application. The user interface was tested by 10 non-skilled users, and the conclusion was that teleDialogue is regarded as being very easy and comfortably to use.

A spin-off effect of this design technique was a wide variety of synonyms for the available services, thereby improving the overall successful transaction rate.

The use of garbage models and models of synonyms for increasing the robustness of teleDialogue has been discussed and considered very important. This goes along with the importance of distribution of the utterances in the database, as well as recording through dialled-up lines. The lack of children in the database was found to lower the overall performance of teleDialogue.

A relatively high number of users are uttering while a message is played back. This indicates that a performance improvement could be expected by including a talk-through facility.

With teleDialogue we have shown a way of integrating speech recognition and user-dialogues into a reliable and robust telephone application that people wants to use in practice.

ACKNOWLEDGEMENT

The authors would like to thank Jørn Stern Nielsen for his help on preparing the statistical data and many helpful suggestions.

REFERENCES

[1] "The Integrated Telecommunication Service (ITS) system", Per Rosenbeck, Signal Processing Group (Jydsk Telefon), COST 232 project, October 1991.

[2] "teleDialog - talegenkendelse i telenettet", Bo Baungaard and Jørn Stern Nielsen, Signal Processing Group (Jydsk Telefon), Tele Teknik, February 1992.

[3] "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", Jay G. Wilpon et al, IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 38, no. 11, November 1990.

[4] "ESPRIT Project 2589 (SAM), SamPac Version 3.10 Documentation, draft", Working Group: Input Assessment, Doc. no. SAM-CT-120, March 1992.

[5] "ESPRIT Project 2589 (SAM), SAM_SCOR Version 3.1 Reference Guide", Working Group: Input Assessment, Doc. no. SAM-IES-066, March 1992.