The CTT-bank demonstration system has been created at the Centre for Speech Technology (CTT) at KTH with the purpose of displaying and evaluating speech technology in telephone banking. When calling the system, users say their name and a short digit sequence of their own choice to identify themselves. Speaker verification with a prompted four-digit password is then used to authenticate identity claims. Dialogues are finite state-based. The system asks for a single piece of information at a time, such as an account name or a money amount, and users may respond with keywords optionally embedded in carrier phrases. Three banking services have been implemented: account balance, list of recent transactions and transfer between accounts. Help is always available upon request and users may interrupt the dialogue at any time to go back to the starting point of a service loop. A small user trial with 24 subjects has been performed. Subjects enrolled in the bank and called the service up to seven times during a two-week period. A questionnaire was used to collect user opinions, and speech data was recorded and used for the evaluation of speech recognition and speaker verification components in the system. The average word error rate for speech recognition of utterances with four to seven digits was 4.6%. Concept error rates for recognition of service commands, account names, money amounts and answers to yes/no questions were 12% on average for utterances where all spoken words were represented in the lexicon. The false reject rate for the speaker verification system was 4.7% with 30 seconds of enrolment data and up to three attempts at the 2-second test utterance. Impostor attempts were not included in the trial and therefore no direct measurements of false accept rates can be made. Previous work with user trials on speaker verification in telephone banking contexts has also been reviewed. Based on additional results in those trials, it is estimated that speaker verification error rates around 1% is achievable in a telephone banking application. Combined with other means of authentication this offers a large increase in security. User opinions about the trial were positive, and the overall conclusion is that speech technology is useful in telephone banking.

1. Introduction

The CTT-bank project started in January 1999 as a co-operation between Trio, Handelsbanken, Ericsson and KTH, within CTT, the Centre for Speech Technology at KTH. At a later stage, the Swedish Handicap Institute (Hjälpmedelsinstitutet) was also partly involved.

The aim of the project was to develop a speech controlled telephone banking demonstration system in order to investigate whether speech technology is useful for telephone banking, and also to develop a dialogue that works well for users. In addition to navigating the banking services verbally, customers were to be authenticated by voice using automatic speaker verification (ASV) technology. Within the project, a study has been performed with 24 subjects calling the banking demonstration system up to seven times during two weeks. No connections to real bank accounts or real money were made in the study.

This paper contains a description of the demonstration system and the user trial together with evaluations of usability and performance of speech technology components. Related work is presented in companion papers: the software platform is further described in Melin (2001), and a more detailed evaluation of the speaker verification subsystem in Melin (to appear). A detailed analysis of usability aspects is presented in Ihse (2000), part of which is reviewed in this paper.
Similar investigations with focus on speaker verification have been conducted in the EU-funded CAVE, PICASSO and SPOTLIGHT projects. Within CAVE, Ubilab, the research laboratory of Swiss UBS, ran a field test with a telephone banking service. It was based on fictitious accounts and involved 185 subjects speaking Swiss German (CAVE consortium, 1998). Within the follow-up project, PICASSO, two field tests related to home banking were made. In the United Kingdom, the Nationwide Building Society (NBS) attached speaker verification as a front-end to their existing call center-based home banking service. 428 customers enrolled into the speaker verification system during a seven-month test period (Coutts, 2000). The Dutch Fortis Bank ran an investment game where participants could obtain the latest stock information and trade stocks (Gu et al., 2000). SPOTLIGHT ran a large scale evaluation of a commercial speaker verifier (Nuance) with close to 1000 speakers of British English in the context of a banking application (Spotlight consortium, 2000). This study focused on the technical performance of speaker verification and did not include issues such as usability and dialogue design. Together with the banking applications, the CAVE and PICASSO projects ran field tests on calling card services (den Os et al., 1997; Moser et al., 1998) and used speaker verification to grant free access to directory assistance (den Os et al., 1999). Boves & den Os (1998) and Boves et al. (1999) discuss several technical and human factors issues related to the use of speaker verification in real-world applications, based on experiences from these two projects.

Several other experiments have been made in different or similar application areas, for example automatic directory assistance services (Billi et al., 1998; Kellner et al., 1998), voice activated dialing services (Tan et al., 1998) and train timetable information systems (Sanderman et al., 1998). Some of them were conducted entirely within research projects, others were open to the public but part of research projects, and still others were commercial applications.

2. System description

From the users’ point of view, the CTT-bank demonstration system was similar to the commercial automated telephone banks prevalent in Sweden, with the difference that all user input to CTT-bank was spoken, as opposed to the commercial systems, where the buttons on the telephone are used (DTMF signaling). The dialogue in the CTT-bank demonstration system was strictly menu-driven, as in the commercial DTMF systems, but the arrangement of the experiment (for example the information to the test subjects and the spoken prompts from the system) was intended to stimulate a more free input from the users than the system could actually handle, in order to collect data for further development towards a more free dialogue.

The services available in the CTT-bank demonstration system were some of the most commonly offered services in commercial telephone banks: statement of balance, information on recent transactions and transfer of funds between the user’s own accounts. The menu hierarchy in CTT-bank imitated that of Handelsbanken’s automated telephone bank, provided by Objecta Systems AB, a subsidiary company of Trio.

2.1. Dialogue design

2.1.1. Authentication

In commercial DTMF-driven telephone banking applications, the login process typically consists of giving a user number (in Sweden often the personal identity number, “personnummer”) and a PIN (personal identification number), often four digits. Speech technology offers other possibilities for claiming identity, primarily through combinations of speech recognition and speaker verification. In CTT-bank the users identified themselves by saying their name followed by a short digit sequence (2-4 digits). The digit sequence was chosen by the user during an enrolment call and had to be unique among users with the same name. They verified the identity claim with the voice by repeating a prompted sequence of four digits.

2.1.2. Registration

Speaker verification can only be accomplished if the system has some prior knowledge of the user’s voice. Therefore the CTT-bank users had to make an enrolment call before they could start using the actual application. To make sure the right person was enrolled, each user had to speak an individual, seven-digit registration number printed on an enrolment form. They then read ten sequences of five digits from the same form. The digit sequences were carefully chosen, so that each digit occurred in different contexts of other digits, to include coarticulation effects in the training material for the speaker...
verification system. At the end of the call the users had to choose the number to use together with their name as a means of identification in the following calls. The system used explicit confirmation to make sure it had recognised the correct number, i.e. the system asked the user “4 7 1 1, is that correct?”.

2.1.3. Banking services

A call to the CTT-bank demonstration system started with a login phase and continued into a service loop that started with the general question “What do you want to do?”. Behind this question were the three available bank services and options for help and for leaving the system. The application restarted from this general question when a service had been fully accomplished or cancelled, as illustrated in Figure 1.

The balance and recent transactions services were straightforward: the user was asked for an account name and was given information about that account. The account represents the single slot of information in these two services. The transfer service had three slots to fill: from which account, to which account and how much money, and each slot had its own question. It may have been more convenient for the user to answer all three questions in one utterance, but in this first version of CTT-bank we chose not to challenge the speech recognition and error handling too much, and therefore we stuck to the strict menu choices of the corresponding commercial application. An implicit confirmation strategy was used with the balance and recent transactions services, while explicit confirmation was used with the transfer service. In the latter case, a single confirmation question was asked after the system had an entry for each of the three slots: “You want to transfer <amount> from <from-account> to <to-account>. Is this correct?”. In the event of a negative response, the system fell back on confirming each slot entry individually.

Some user commands were always available: help, cancel and quit. The help service was context-sensitive. It presented a help message that was different depending on where in the dialogue help was asked for. It was also possible to get additional help after the first help message had been presented. The cancel command interrupted what was going on and brought the

---

4 For example, if users asked for the balance of their savings account, the system responded “the balance of your savings account is ...”.

Figure 1. The service loop in a CTT-bank call.
dialogue to the starting point of the service loop. Quit resulted in a short goodbye message and a terminated call. An application used in real life would have given the possibility of hanging up anywhere in the dialogue, but the subjects in this study were instructed to end the call by explicitly saying “quit”, mainly because the software was by then not robust enough to handle calls ended by hanging up the phone.

2.2. Implementation framework

The CTT-bank system is built on top of ATLAS, a generic Java software library that provides a framework for building multi-lingual and multi-modal applications on top of speech technology components (Melin, 2001). The framework provides a powerful speech technology application programming interface (API) to ATLAS and the underlying components, with the aim of collecting as much of application-independent functionality as possible within the framework. A schematic view of the system is shown in Figure 2, with the CTT-bank application communicating with a set of components through ATLAS. The following sections describe the application-dependent CTT-bank layer and the main speech technology components in the system: the speech recognition (ASR), speaker verification (ASV), text-to-speech (TTS), and speech detector engines.

2.3. The application

The application-dependent layer in CTT-bank defines several dialogue components to implement the banking services and part of the enrolment dialogue. They all derive from the same component called “complex question”, and their respective task is to get a money amount, to get the name of a valid account, and to get the answer to a yes/no-question. Dialogue components use various ATLAS layers for their implementation.

ATLAS own dialogue components for enrolment and login are extended and specialised, and used to implement enrolment and user authentication dialogues in CTT-bank. Specialisation includes using an error-correcting code with the seven-digit registration number used to authenticate the user during the enrolment call, and changing prompt texts to fit the application.

2.4. The speech recognition system

The speech recognition component is built on the StarLite recognition engine (Strom, 1996). StarLite was developed during the WAXHOLM project, where it was used for a medium-size vocabulary recognition task. It has later also been used in the Gulan (Stjólander & Gustafson, 1997) and the August systems (Gustafson et al., 1999). CTT-bank is the first application where StarLite has been used with telephone quality

![Figure 2. The system model used in the CTT-bank system. It is layered with an application-dependent layer on top, a resource layer at the bottom, and an application-independent layer (ATLAS) in between.](image-url)
speech input. This section describes the setup of the speech recognition system as used in CTT-bank, with emphasis on the acoustic and language models.

### 2.4.1. Feature extraction

The input signal is pre-emphasised and divided into one 32 ms frame each 10 ms and a Hamming window is applied. For each frame a 12-element cepstral vector and an energy term is computed, and they are appended with first and second order deltas. Cepstral vectors are computed from a 24-channel, FFT-based, mel-warped, log-amplitude filterbank between 200-3800 Hz followed by a cosine transform and cepstral liftering. The energy term is the 0\textsuperscript{th} cepstral coefficient. The total vector dimension is 39. Cepstral mean normalisation is not used.

### 2.4.2. Acoustic models

The acoustic models were created by Salvi (1998). They are context-dependent, phone-level, left-to-right, tied-mixture hidden Markov models (HMM) trained on 800 speakers (461 female, 339 male) drawn from the 1000-speaker version of the Swedish SpeechDat database with recordings from the fixed telephone network (Elenius, 2000; Höge et al., 1997). In the subset of the database used in training these models, speakers read sentences (nine per speaker) and phonetically rich words (four per speaker) drawn from the text of newspaper articles. The same model set is used in all recognition tasks in the application, each corresponding to a grammar included in Table 1. Thus, the acoustic models are trained for domain-neutral, read speech, rather than domain-specific, spontaneous speech as is in fact spoken in the limited domain of a telephone banking application. This arrangement was motivated for our experiment by the simplicity in setting up the recogniser for each new task. It can be expected, however, that the use of domain-dependent acoustic models, especially word models, would increase recognition accuracy, but at a higher development cost.

Each phone HMM has three states (plosives four states), and the observation probability density function (pdf) associated with each state is a four-component Gaussian mixture with diagonal covariance matrix. To ensure that model parameters were robustly estimated during training, HMM states were tied across the model set using a top-down phonetic decision tree clustering. The aim of clustering is to balance the number of parameters to estimate, and their distribution across models, against the available amount of training data. Clustering resulted in an HMM set with a total of 2020 unique states with approximately 4.5 hours of training data.

### 2.4.3. Language models

The StarLite engine can use either of two types of language models; class bigram-grammars or class pair-grammars. In the CTT-bank application the class pair-grammar alternative is used to implement a set of small finite state grammars, one for each type of utterance the system expects from the user. To simplify

---

**Table 1.** A list of the eight grammars used in speech recognition in the CTT-bank application. The vocabulary size column shows the number of concepts included in the grammar for amount, account and service grammars, and regular vocabulary sizes for the other grammars. The number of additional words or lexicalised phrases allowed in carrier phrases are shown within parentheses. Words belonging to the carrier phrase are shown in italics in the example utterance column. The interrupt grammar is always used in parallel to each of the amount, account and service grammars.

<table>
<thead>
<tr>
<th>grammar</th>
<th>type of utterance</th>
<th>vocab. size</th>
<th>example utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 digits</td>
<td>a prompted password string</td>
<td>10</td>
<td>3520</td>
</tr>
<tr>
<td>5 digits</td>
<td>an enrolment utterance</td>
<td>10</td>
<td>14472</td>
</tr>
<tr>
<td>7 digits</td>
<td>a registration code used for authentication in the enrolment call</td>
<td>10</td>
<td>4264589</td>
</tr>
<tr>
<td>amount</td>
<td>a currency amount</td>
<td>30 (+1)</td>
<td>två tusen fem hundra kronor (2500 kr)</td>
</tr>
<tr>
<td>account</td>
<td>the name of a bank account</td>
<td>8 (+7)</td>
<td>mitt sparkonto (my savings account)</td>
</tr>
<tr>
<td>service</td>
<td>the request for a banking service</td>
<td>10 (+11)</td>
<td>jag vill göra en överföring (I want to make a transfer)</td>
</tr>
<tr>
<td>interrupt</td>
<td>help, abort and quit commands</td>
<td>3 (+4)</td>
<td>jag behöver hjälp (I need help)</td>
</tr>
<tr>
<td>yesno</td>
<td>the answer to a yes/no question</td>
<td>4 (+3)</td>
<td>ja (yes)</td>
</tr>
</tbody>
</table>
subsequent processing (language understanding), the grammars are designed to match keywords in an optional carrier phrase. For example, when a client is asked for a target account for a money transfer, the expected keywords are the predefined names of accounts (salary account, savings account, etc.). Allowed carrier phrases include “to the <account>” or “my <account>”. When asked what service track the client wants to take, keywords are names of services (check balance, transfer funds, etc.) and carrier phrases include “I want to <service>”. All synonyms of a keyword are coded with the keyword itself as the output symbol, and words in carrier phrases are coded with a null output symbol. The eight grammars used in the application are listed in Table 1. A finite state grammar is implemented with a class-pair-grammar by defining unique classes for each state as illustrated in Figure 3.

2.4.4. Decoding
Decoding is performed in a two-pass search: the first pass is a Viterbi beam-search and the second is an A* stack decoding search. The output from decoding is an n-best list of multiple hypotheses for the text of the utterance (including word boundary timing information used by the speaker verification component).

2.5. The speaker verification system
The speaker verification component in the CTT-bank system is built on GIVES, a generic platform for speaker verification systems developed at KTH/TMH. The system is text-dependent and operates in a text-prompted mode with digit string utterances. Except for two modifications, it is the same as the baseline system described and tested in Melin et al. [1998] and Melin & Lindberg (1999). This section reviews the design of the system, and indicates with footnotes where the system differs from the baseline system.

2.5.1. Feature extraction
The input signal is pre-emphasised and divided into one 25.6 ms frame each 10 ms and a Hamming window is applied. For each frame a 12-element cepstral vector and an energy term is computed, and they are appended with first and second order deltas. Cepstral mean subtraction is applied to the 13 basic coefficients. Cepstral vectors are computed from a 24-channel, FFT-based, mel-warped, log-amplitude filter bank between 300-3400 Hz followed by a cosine transform and cepstral liftering. The energy term is the 0’th cepstral coefficient. The total vector dimension is 39. Note that this processing differs from the feature extraction for the speech recogniser in the previous section only in the window length, the filter bank frequency range, and the use of cepstral mean subtraction (CMS). The more narrow frequency range and the use of CMS in the speaker verifier are intended for robustness against inter-session channel variation.

2.5.2. Speaker models
Two kinds of speaker models are used in the system: client models that represent the voices of particular speakers (CTT-bank customers), and non-client models that represent the voices of universal groups of speakers. While client

![Figure 3. a) class-pair grammar for a four-digit string, and b) the equivalent state diagram. States in the diagram correspond to classes in the grammar.](image-url)
models are single-speaker models, non-client models are multi-speaker models. The system has two gender-dependent non-client models: one trained on a large population of male speakers and the other on a large population of female speakers. Non-client models are used for two purposes: as seed models during enrolment as described in this section, and for score normalisation as described in the following section.

Both kinds of speaker models have 10 word-level left-to-right hidden Markov models (HMM), one for each digit. Each HMM has two states per phoneme in the modelled word, and the observation probability density function (pdf) associated with each state is an eight-component Gaussian mixture with diagonal covariance matrix. Each HMM in a client model is trained separately from other HMMs in the model. It is thereby assumed that all relevant words and their boundary locations in enrolment utterances are known.

Rather than training a client model directly from enrolment data, we use the following adaptation procedure. For each word model in the client model, a corresponding word model from one of the two non-client models is selected as a seed model based on their closeness to enrolment data. That is, if the male model fits better to the data, the male model is chosen, otherwise the female model is chosen. Note that no a priori information about the gender of the client is used in this selection, the selection is made on a word-by-word basis, and there is no restriction that all word models must be selected from the same gender model.

The seed model is then used as the basis for client model adaptation: transition probabilities and variance vectors are left as they are, while mean vectors and mixture weights are trained from the collected data. Training is performed with the Expectation Maximisation (EM) algorithm to optimise the Maximum Likelihood criterion

\[
\left( c^*_w, m^*_w \right) = \arg \max_{(c, m)} P(O_w | c_w, m_w, \sigma^2_w, A_w),
\]

where \( O_w \) is the enrolment data for word \( w \), \( c_w \) and \( m_w \) are vectors of all mixture weights and mean values in the HMM to optimise, and \( \sigma^2_w \) and \( A_w \) are the fixed variance and transition probabilities taken verbatim from the seed HMM. The seed means and mixture weights are used as the starting guess in the first iteration of the EM algorithm.

2.5.3. Classification

A verification test is based on a log-likelihood ratio test over each word segment in the test utterance. The log-likelihood ratio score is normalised to the length of the word, and such normalised scores are averaged over all words in the utterance to produce the decision variable. The selection of a background model for each of the log-likelihood ratio tests is based on the similarity between each of the two (male/female) non-client word models and the test segment\(^5\). As during enrolment, the selection is made on a word-by-word basis.

The classifier decision is taken by comparing the value of the decision variable to a speaker-independent threshold. The value of the threshold was set to the same-sex equal error rate (EER) threshold in a calibration experiment on the Gandalf database [Melin, to appear].

2.5.4. Segmentation

As indicated above, the system depends upon explicit segmentation of input speech into words during both enrolment and test. In the current setup, this segmentation is taken from the speech recognition component of the CTT-bank system. The speech recogniser engine produces a list of hypotheses for a spoken utterance, and each hypothesis is defined by a text, a score value and a word-level segmentation of the utterance. A hypothesis is selected based on the prior knowledge of what the user is supposed to say, as the hypothesis with the highest score, the text of which is identical to the expected text\(^6\). (Note that if there is no such hypothesis, the dialogue system will have rejected the utterance and prompted the user for a new one.) The system always knows what text to expect from the user, since pre-defined digit strings are used during enrolment, and the password phrase used during a test is generated by the system and given to the user through an audio prompt.

2.6. The text-to-speech system

The text-to-speech (TTS) system consists of a rule-based text-to-phone converter [Carlson et al., 1982] and an MBROLA synthesiser [Dutoit et al., 1996] with the Lukas voice [Filipson & & Lindberg (1999)], where selection was based on similarity to enrolment data rather than test data.

\(^5\) The first difference to the baseline system in Melin & Lindberg (1999), where selection was based on similarity to enrolment data rather than test data.

\(^6\) The second difference: in Melin & Lindberg (1999), segmentation was created via forced alignment to what the user was supposed to say, regardless of what he actually said. The acoustic models were also different.
The speech signal is generated with 16 kHz sampling frequency and is downsampled to 8 kHz in the telephony server using linear interpolation. In several cases, prompt texts sent to the TTS were hand tuned before the user trial for improved TTS output quality. Hand tuning included adding word stress markers and transcribing words that were otherwise mispronounced.

2.7. The speech detector

The speech detector component inputs the recorded signal, and decides when the user starts and stops speaking, before sending the signal to the speech recogniser. It uses energy and zero-crossing rate in the speech signal.

The speech detector was written in C++ by Bruce T. Lowerre⁷, but has been ported to Java at KTH/CTT and adapted to ATLAS. A time-out feature for complete silence has been added so the application can react when the user does not respond to a question within a given time.

3. Experiments

Data were collected in two steps. First, a small preliminary study was made to provide input for system design. Once the system had been designed and implemented, the main study was conducted. This section describes both studies.

3.1. Preliminary study

Four volunteers who were used to telephone banking services were asked questions about their use. The questions covered how frequently they called the bank; what type of operations they performed; if they used the telephone bank for regular and recurring operations or just on special occasions; if they usually did more than one operation per call; if they used the telephone bank (as opposed to other methods, such as Internet banking) under special conditions; and if there were any peculiarities with their normal telephone banks they were discontent with.

The answers indicated that the subjects called their telephone banks several times per month, more often at the end of the month. (Note that the unscientific selection of subjects matters here: only people who were regular telephone bank users were asked to take part.) The most common operations were checking the balance and transferring funds between the subject’s own accounts. It was also suggested that the combination of the two is very common, and that other combinations are common as well.

Telephone banking was used primarily as a complement to Internet banking when access to the Internet was not practical. Three of the four subjects used telephone banking especially from their mobile phones. As for peculiarities, some examples were reported that could serve as good ideas of what not to do in CTT-bank:

- not offering a list of available accounts
- requiring the user to enter account numbers
- recorded advertisement for the bank’s services
- everything that hindered rapid service completion by the advanced user, such as not allowing barge-in on system prompts.

As a result of the preliminary study, several design decisions were made. These included making the dialogues as fast as possible for the experienced user, telling the balance of the account before making a transfer from it, and assuming that the user normally wanted to perform several operations during one call.

3.2. Subjects

For the main study, 28 subjects were recruited. 24 were chosen by CTT’s industrial partners involved in CTT-bank (15 from Handelsbanken, 5 from Trio and 4 from the Swedish Handicap Institute), whereas 4 were KTH students.

It should be pointed out that the selection of subjects was in no way random, and that this most likely influenced the study. First and foremost, the group of subjects provided by the Swedish Handicap Institute (HI) deliberately deviated from a statistically representative selection of the general population. They were all disabled, as a result from an explicit wish from HI to examine the possible impact of speech technology on disabled people. But hopefully the 28 subjects represent different groups with different prior knowledge of telephone banking and speech recognition systems, making the results not too skewed. It is our understanding that the Handelsbanken subjects (all bank employees) were accustomed to telephone banking, but perhaps not to speech technology. The Trio subjects (all employees at Trio) were more familiar with technology in general, and with automatic telephone systems in particular. The KTH subjects, finally, are probably representative both regarding speech technology and telephone banking.

Subjects were told that calls to CTT-bank would be recorded and saved for future research. Since the login procedure included saying their names, they were given the option to choose pseudonyms for login names.

⁷ ftp://svr-flp.eng.cam.ac.uk/pub/comp.speech/tools/
3.3. Realization of the study

Before the study, all subjects were informed of its goals and their expected part in it. To make sure all subjects got the same information, they were all sent identical documents: One described the study and explained the receiver’s role as a subject (the Study information paper). The other document was written to look like an official paper coming from a real bank (the Bank information paper). The two documents were included in full as appendices in Ihse (2000).

In particular, the Study information paper asked the subjects to call CTT-bank at least three times, preferably five, and at most seven times during a two week period. The subjects were asked to call CTT-bank immediately after calling their normal telephone bank, if they used one regularly, and try to mimic the same operations. If this was not possible, because they would not use a real telephone bank at least three times during the two week period, the subjects were instead asked to try CTT-bank from different phones (mobile phones, home and work phones) and at different times of the day.

The Bank information paper explained the enrolment and the login procedure in detail, and explained only briefly how to use the actual banking services. It included the seven-digit registration code needed for authentication during the enrolment call.

To prepare the CTT-bank system for the study, names of the 28 subjects were entered into the system database along with their assigned seven-digit registration codes. Phonetic transcriptions of the names were included in the speech recognition grammar.

The study started on June 7, 2000 and ended two weeks later. During this time, the subjects were requested to first enrol, and then to call CTT-bank for banking services. The system counted how many logins each subject made, and blocked the account after seven successful logins. This limit was included to make sure no subject called an unreasonable number of times. By limiting the number of allowed calls, it was also hopefully made clear to the subject that each call was important.

To facilitate the analysis of entire conversations, all telephone calls to the CTT-bank system were bridged through Trio Televoice equipment as illustrated in Figure 4. The equipment was set up to record the two sides of the conversation in separate files, whereas the CTT-bank system itself recorded only the user side of the conversation, and only at times when the system accepted input from the user.

3.4. Follow-up

After the end of the main study, the subjects were asked to evaluate their experience of CTT-bank in a questionnaire. The questionnaire was not anonymous, to enable matching users’ subjective experience of CTT-bank with different kinds of objective measurements. The complete questionnaire was included as an appendix in Ihse (2000).

4. Database

Out of the 28 recruited subjects (17 male, 11 female), 24 (13 male, 11 female) made at least one attempt to call CTT-bank, and 21 succeeded with the enrolment call. Among the 21 subjects who completed the enrolment call were 11 men and 10 women. Their age distribution is shown in Figure 5. All subjects lived in the Stockholm area.
Table 2. Statistics on the recorded speech files over prompt type. The all-row includes all files, while remaining rows group files according to the type of utterance the system expected from the user. Columns have the following meanings: subjects - the number of users represented in the group, sessions - number of complete calls / number of failed calls, files - number of files, words - number of unique word forms (vocabulary size), phrases - number of unique phrases, empty - number of files with no spoken words. Data in the three last columns are based on orthographic transcriptions.

<table>
<thead>
<tr>
<th>subgroup</th>
<th>code</th>
<th>subjects</th>
<th>sessions</th>
<th>files</th>
<th>words</th>
<th>phrases</th>
<th>empty</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>all</td>
<td>24</td>
<td>82 / 30</td>
<td>1518</td>
<td>214</td>
<td>364</td>
<td>44</td>
</tr>
<tr>
<td>7 digits d7</td>
<td>24</td>
<td>21 / 19</td>
<td>48</td>
<td>10</td>
<td>30</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5 digits d5</td>
<td>21</td>
<td>21 / 8</td>
<td>280</td>
<td>11</td>
<td>13</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4 digits r4</td>
<td>20</td>
<td>61 / 3</td>
<td>85</td>
<td>11</td>
<td>83</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ID-number nm</td>
<td>21</td>
<td>21 / 1</td>
<td>32</td>
<td>10</td>
<td>21</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>name+ID-number id</td>
<td>20</td>
<td>61 / 11</td>
<td>119</td>
<td>52</td>
<td>30</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>amount am</td>
<td>18</td>
<td>40 / 0</td>
<td>73</td>
<td>34</td>
<td>47</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>yes/no yn</td>
<td>21</td>
<td>70 / 1</td>
<td>231</td>
<td>34</td>
<td>18</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>service se</td>
<td>20</td>
<td>61 / 0</td>
<td>305</td>
<td>81</td>
<td>79</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>account ac</td>
<td>20</td>
<td>58 / 0</td>
<td>342</td>
<td>75</td>
<td>66</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>message ms</td>
<td>3</td>
<td>3 / 0</td>
<td>3</td>
<td>44</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 to Table 4 present statistics on the recorded files over various file subsets. In Table 2, files have been grouped according to the type of prompt they were recorded in response to, and in Table 3 according to the subject who is speaking. Table 4 groups files according to call type, where call types are complete enrolment and complete login on the one hand, and failed enrolment and failed login on the other. A complete enrolment call is one where the entire enrolment procedure was completed, while a complete login call is one where the caller managed to log in to the service, that is to successfully pass both the identification and verification stages.

Except for the code column in Table 2, all three tables have the same columns. The subjects, sessions, and files columns contain counts of how many subjects, sessions and files are represented in a subgroup. For example, Table 2 shows that there are 305 files recorded in response to the question what service track a user wanted to take; 20 different speakers are represented in this group of files; and the files were recorded during 61 different (complete) sessions. The sessions column includes two numbers: the number of sessions classified as complete and failed, respectively. Data in the three remaining columns are based on transcriptions of each of the recorded files. The words column shows the number of unique word forms that were used, that is the vocabulary size, while the phrases column shows the number of unique phrases that were used. The empty column, finally, shows the number of files in a subgroup where no words were spoken. In the service example, 81 different words were used in the responses, composed into 79 different phrases. In 10 of the files the user did not say anything.

Speech files recorded by the CTT-bank system were transcribed orthographically using SpeechDat conventions (Senia & van Velden, 1997) for labelling speaker noise, filled pause, stationary noise, intermittent noise, truncated utterances, and misspronounced words. These manual transcriptions have later been used as reference when computing speech recognition error rates.
Table 3. Statistics on the recorded speech files over subjects. The all-row includes all files, while remaining rows group files according to user. Column meanings are described in text by Table 2. The first column shows a code assigned to each of the subjects in the user trial. Codes for female subjects start with an f and those for male subjects start with an m.

<table>
<thead>
<tr>
<th>subgroup</th>
<th>subjects</th>
<th>sessions</th>
<th>files</th>
<th>words</th>
<th>phrases</th>
<th>empty</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>24</td>
<td>82 / 30</td>
<td>1518</td>
<td>214</td>
<td>364</td>
<td>44</td>
</tr>
<tr>
<td>f001</td>
<td>1</td>
<td>3 / 0</td>
<td>41</td>
<td>25</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>f002</td>
<td>1</td>
<td>5 / 1</td>
<td>137</td>
<td>46</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>f003</td>
<td>1</td>
<td>4 / 0</td>
<td>79</td>
<td>46</td>
<td>43</td>
<td>2</td>
</tr>
<tr>
<td>f004</td>
<td>1</td>
<td>5 / 2</td>
<td>82</td>
<td>35</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>f005</td>
<td>1</td>
<td>4 / 0</td>
<td>45</td>
<td>29</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>f006</td>
<td>1</td>
<td>5 / 2</td>
<td>68</td>
<td>43</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>f007</td>
<td>1</td>
<td>1 / 5</td>
<td>39</td>
<td>11</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>f008</td>
<td>1</td>
<td>3 / 4</td>
<td>51</td>
<td>28</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>f009</td>
<td>1</td>
<td>3 / 1</td>
<td>45</td>
<td>48</td>
<td>33</td>
<td>0</td>
</tr>
<tr>
<td>f010</td>
<td>1</td>
<td>3 / 0</td>
<td>41</td>
<td>42</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>f011</td>
<td>1</td>
<td>0 / 2</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m001</td>
<td>1</td>
<td>5 / 1</td>
<td>164</td>
<td>43</td>
<td>55</td>
<td>15</td>
</tr>
<tr>
<td>m002</td>
<td>1</td>
<td>4 / 0</td>
<td>36</td>
<td>27</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>m003</td>
<td>1</td>
<td>5 / 1</td>
<td>74</td>
<td>99</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>m004</td>
<td>1</td>
<td>4 / 1</td>
<td>107</td>
<td>53</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>m005</td>
<td>1</td>
<td>6 / 1</td>
<td>121</td>
<td>54</td>
<td>51</td>
<td>3</td>
</tr>
<tr>
<td>m006</td>
<td>1</td>
<td>3 / 0</td>
<td>57</td>
<td>45</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>m007</td>
<td>1</td>
<td>7 / 0</td>
<td>129</td>
<td>70</td>
<td>66</td>
<td>1</td>
</tr>
<tr>
<td>m008</td>
<td>1</td>
<td>4 / 0</td>
<td>56</td>
<td>46</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>m009</td>
<td>1</td>
<td>2 / 0</td>
<td>44</td>
<td>32</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>m010</td>
<td>1</td>
<td>3 / 0</td>
<td>42</td>
<td>25</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>m011</td>
<td>1</td>
<td>3 / 5</td>
<td>54</td>
<td>31</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>m012</td>
<td>1</td>
<td>0 / 3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>m013</td>
<td>1</td>
<td>0 / 1</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4. Statistics on the recorded speech files over call types. The all-row includes all files, while remaining rows group files according to call type. Column meanings are described in text by Table 2. A failed session is one that was ended prematurely either by the subject hanging up or the system crashing.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>subjects</th>
<th>sessions</th>
<th>files</th>
<th>words</th>
<th>phrases</th>
<th>empty</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>24</td>
<td>82 / 30</td>
<td>1518</td>
<td>214</td>
<td>364</td>
<td>44</td>
</tr>
<tr>
<td>complete enrolments</td>
<td>21</td>
<td>21 / 0</td>
<td>350</td>
<td>12</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>failed enrolments</td>
<td>10</td>
<td>0 / 19</td>
<td>67</td>
<td>11</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td>complete logins</td>
<td>20</td>
<td>61 / 0</td>
<td>1058</td>
<td>214</td>
<td>294</td>
<td>33</td>
</tr>
<tr>
<td>failed logins</td>
<td>7</td>
<td>0 / 11</td>
<td>43</td>
<td>24</td>
<td>19</td>
<td>5</td>
</tr>
</tbody>
</table>
5. Results

Of the calls reaching the bridging point, not all reached the application. The application hard-
ware and software did not handle simultaneous calls and therefore the line was sometimes busy.
The enrolment calls and the bank service calls had two different telephone numbers and could therefore easily be classified even when they did not reach the application.

In total, 237 calls reached the bridging point and were recorded there. Of these, 102 were calls to the enrolment line and 135 to the bank service line. Only 112 of the 237 calls reached the application (40 for the enrolment and 72 for the bank service). For the remaining calls, the line was busy in 71 cases, system errors were at hand in at least 10 cases and the rest were due to other problems, for example users hanging up before the dialogue started.

5.1. Speech recognition

Speech recognition results are presented across utterance types and across speakers in terms of error rates and confusion matrices. Word error rates (WER) are used with utterances such as digit strings where all words are important for the application and few out-of-vocabulary (OOV) words are expected. For utterances where only keywords are important and many OOV words may be expected, concept error rates (CER) are used instead.

Two types of WER and CER test sets are used in this paper. The first, the first attempt test set, includes the first recorded attempt on each slot within each utterance group (where a group is defined by an utterance type or a speaker), while the second, the all attempts test set, includes all recorded attempts. Results on the first attempt test set are better suited for comparison with results from tests on speech corpora such as SpeechDat, recorded by a passive dialogue system and with no system feedback. In such corpora, the available data may be regarded as “the first attempt”. Results on the all attempts test set, on the other hand, summarises how many errors the system made in total, given that one error may have generated a new attempt on the same input slot.

To fully appreciate the difference between the two test sets, consider how attempts are generated. Firstly, the system usually has some way of confirming that the correct words were recognised, through prior knowledge or implicit or explicit confirmation. When the system determines that the recognised text is not likely to be correct, it asks the user to try again. The result is that for each slot of spoken user input, between one and $N$ attempts will be recorded, where $N$ is the maximum number of attempts the system allows for ($N$ depends on the utterance type). Most of the last recorded attempts were accepted by the system, while most attempts before the last one were rejected. Including all attempts in a test set, rather than selecting a single attempt per item, may bias the result to emphasise recogniser mistakes. For example, if the recogniser makes the same mistake in two consecutive attempts, that mistake will be included twice in the result. Secondly, speakers for whom a recogniser makes many mistakes (often called goats) will generate more attempts per slot with recognition errors than speakers for whom the recogniser works well (often called sheep).

The first attempt test set is easily defined for the speech material consisting of names and digit strings, as these are always explicitly prompted by the application. For answers to normal questions, on the other hand, the definition of a first attempt is more difficult to make. In the question-answer dialogues, the system asks the user to make a choice, and the application therefore knows less about what utterances to expect. For example, if the system asks for an account, the user may respond with the name of an account, or may instead choose to ask for help or to cancel the current operation. For names and digits, the system also has ways to confirm that the correct words are recognised, and the main cause for extra attempts is recognition errors. For question-answers, on the other hand, extra attempts may result from the users changing their minds. Another example of a second attempt that could be classified as a first attempt, is when the user’s first answer is a request for help, and the second answer is the choice after help has been given.

To avoid subjective judgement of what is a first attempt and what is not, we have chosen to strictly adhere to the definition of a slot as any number of consequent repetitions of the same question, disregarding any intervening non-question prompts (for instance help messages). The above definition of the first attempt test set is then still valid: it includes the answer to each of all slot-initial instances of the same question. The all attempts test set for a given utterance

---

$^8$ The last attempt may not have been accepted if the call was ended after it was recorded, either because the maximum number of attempts were used, the system crashed or the user hung up.

$^9$ All attempts except for prompted passwords where an utterance may have been rejected by the speaker verifier rather than the speech recogniser.
type includes the answer to all questions in slots corresponding to that utterance type.

All presented results are based on offline simulations with audio files recorded during the user trial. They differ from actual results produced during the trial in the following respects:

- There is a recognition result for every recorded file. In some cases during the user trial the CTT-bank system crashed in such a way that a file was recorded but no recognition result was produced.
- Faulty phonetic transcriptions of the word “9” in digit string lexicon files have been corrected. During the user trial, transcriptions were (using the International Phonetic Alphabet, IPA)

<table>
<thead>
<tr>
<th>spoken</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>del.</th>
<th>% error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>168</td>
<td>1</td>
<td>2</td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>4.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>151</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
<td>191</td>
<td>4</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>4.0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td>175</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td>146</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>20</td>
<td>1</td>
<td>128</td>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
<td>16.3</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td>2</td>
<td>159</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>154</td>
<td>0</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>163</td>
<td>0</td>
<td>4.1</td>
</tr>
</tbody>
</table>

| ins.  | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |     |          |

Figure 6. Confusion matrix for digit recognition. Results are based on the top hypothesis on the first attempt on each d7, d5 and r4 slot (first attempt test sets). The % error-column shows the rate at which a spoken word was recognised as something else (it does not include deletion errors). In one file the speaker said “80” instead of “8 0”. This was recognised as “8”. The corresponding row has been omitted from this confusion matrix.

5.1.1. Names and digit strings

Digit strings were used in utterances related to enrolment and login. During each (successful) enrolment call, subjects first spoke a seven-digit number (d7) to identify themselves, then ten five-digit strings (d5) to provide training data for their speaker model, and finally the ID-number (nm) they wished to use. During the initial login phase of each call to the bank service line, subjects first spoke their name plus their chosen ID-number (id), and then repeated a random four-digit password (r4) prompted by the system.

The system had a dedicated grammar model for each of these utterance types. The grammars were designed for full recognition of each utterance, i.e. the grammars did not include additional carrier phrases or OOV words. Transcriptions of recorded utterances also show that indeed very few OOV words were used (see the words column in Table 2). It is therefore
appropriate to evaluate recognition results in terms of word error rates (WER). For id utterances, the given name and surname of each subject were coded in the grammar as one unit. The same coding was used when computing error rates, and therefore the WER is not strictly a word error rate for id utterances.

For each of the five utterance types listed above, the system had a way to confirm that the correct words were recognised. For d5 and r4 utterances, prior knowledge was used. The system already knew what the users were supposed to say; their purpose was for the users to show the system how they said the given words. For the d7 and id utterances, the recogniser output was validated through a database lookup. The text had to match an entry corresponding to a customer identity in the database. In addition to database lookup, an error correcting code able to correct recognition errors in one of the seven digits was used to define valid numbers for d7 utterances. For ID-number utterances (nm), finally, explicit confirmation was used. The system repeated what was recognised and asked the user if it was correct.

For d7, d5, r4 and id utterances, the respective confirmation strategy was also used to select a hypothesis from the \( n \)-best list output by the speech recogniser. The first hypothesis in the list that lead to a positive confirmation was promoted and used in subsequent processing steps.

Table 5 presents word error rates for each of the utterance types with name and/or digit strings. The top columns show error rates for the top hypothesis in the recogniser’s \( n \)-best list (\( n = 10 \)), while word error rates after the above mentioned hypothesis promotion are shown in the best column.

Table 6. Concept error rates in percent (%) for utterances with answers to questions. Italic numbers indicate the number of utterances in each group.

<table>
<thead>
<tr>
<th>utterance type</th>
<th>code</th>
<th>lexical</th>
<th>non-lexical</th>
<th>lexical + non-lexical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>first attempt</td>
<td>all attempts</td>
<td>first attempt</td>
</tr>
<tr>
<td>account</td>
<td>ac</td>
<td>15.6</td>
<td>211</td>
<td>16.4</td>
</tr>
<tr>
<td>amount</td>
<td>am</td>
<td>18.9</td>
<td>53</td>
<td>20.0</td>
</tr>
<tr>
<td>service</td>
<td>se</td>
<td>7.6</td>
<td>171</td>
<td>10.1</td>
</tr>
<tr>
<td>yes/no</td>
<td>yn</td>
<td>10.2</td>
<td>196</td>
<td>10.0</td>
</tr>
<tr>
<td>all types</td>
<td></td>
<td>12.0</td>
<td>13.1</td>
<td>62.3</td>
</tr>
</tbody>
</table>

Figure 6 shows a confusion matrix for digit recognition, based on the first attempt on each d7, d5 and r4 slot pooled together. The total first attempt WER is 4.6% (out of 1668 words). Errors, as computed by the default HTK evaluation algorithm (Young et al., 1999), are distributed as 60 substitutions, 9 deletions and 7 insertions. The corresponding sentence error rate is 18% (out of 337 utterances). A sentence is correctly recognised if there are no word errors. The figure shows that the main confusion is “6” recognised as “1”, accounting for 26% of all the word errors. Half of the 20 6-to-1 confusions occurred for a single (female) speaker for which “6” was consistently recognised as “1”. 17 of the 20 confusions occurred for female speakers.

Looking at the effect of the error correcting code (ECC) used with d7 utterances, it appears that the code was able to correct ASR errors in 6 out of the 7 files (86%) where such an error occurred and the entire utterance was captured by the system. For the 6 files the speech recogniser made a single substitution error in the best hypothesis, while in the 7th there was more than one error in the best hypothesis. There were a total of 48 d7-files. In 39 of them (81%), the entire utterance was captured. In the remaining 9 files (19%), two or more digits were not captured by the system for one of the following reasons: the user spoke before recording started (2 cases), the user said nothing (1 case), the speech detector cut off in the middle of the utterance (4 cases), or another system error (2 cases). These kinds of errors were never the target for ECC.

As noted above, there was a transcription error for the word “9” in the pronunciation dictionaries used with the digit sequence utterances (d7, d5 and r4) during the trial. Recognition results presented above were re-
produced after the error had been corrected. The effect of the correction was that the top hypothesis output changed in 13 of the 413 files (3.1%). In 2 of the 13 cases the change was to the worse and in 6 cases to the better. In the remaining 5 cases, the result was different but equally right or wrong. Hence, the transcription correction improved recognition rate on average, but the difference was small.

5.1.2. Answers to questions

After the login phase of a call, most of the user input is answers to questions from the application, and the dialogue path through the application is not as strict as in the enrolment call or the login phase. The path always starts with the question “What do you want to do?” and from there, the continuation of the call depends on choices from the user. Each question from the application corresponds to one of four utterance types from the user: service (for the question already mentioned), account, amount and yes/no. Each utterance type has its own grammar. There is also a supplementary grammar called interrupt, that consists of phrases for getting help, cancelling an operation or quitting. This grammar is loaded together with each of the other grammars. The recognition output from the answers-to-questions grammars is only keywords. One keyword can correspond to a variety of words and different pronunciations of these, and there are also carrier phrases like “I want to…”,” “to…” and “from…”, that would be ignored by the application and therefore need no output symbol. The evaluation of the recognition of answers to questions is therefore made with concept error rates (CER) instead of the word error rates (WER) used with digits and names.

Our intention with starting with the general question “What do you want to do?” was to invite the users to speak naturally to the application, and this resulted in many out-of-vocabulary (OOV) words and out-of-grammar (OOG) sentences. The recorded speech material of answers to questions has been divided into two subsets: lexical utterances (with full correspondence in the grammars) and non-lexical utterances (which are only partly or not at all represented in the grammars). In many cases, even the non-lexical utterances gave recognition results (keywords) that resulted in a correct response from the application, and the non-lexicality ranges from very close to the grammars to far away from them.

11 Actually, two engines with different grammars are run in parallel.

The service grammar differs from the others in that it allows more than one keyword in the output of the recogniser. After recognition has been performed, the application uses a priority list to choose which keyword to act upon. Lexical utterances most often result in only one keyword, while non-lexical utterances sometimes result in several keywords, in many cases due to erroneous guesses from the recognition engine when an utterance contains many OOV words. Therefore, when the correct keyword is recognised but another erroneously recognised keyword has higher priority, then the whole utterance is misinterpreted, leading to the wrong response from the application.

The evaluation of the recognition of non-lexical utterances is not made in the same detail as for lexical utterances. For example, no confusion matrices are presented. The main reason for calculating recognition error rates of non-lexical utterances is for comparison with the subjects’ experiences of how well they were understood by the system (Figure 10 and section 5.4.1).

Several experiences can be drawn from how the grammars were made:

- Account names not used in the application or by the specific user should not be included in the grammar. In this study, all users had the same set of accounts: savings account (sparkonto), salary account (lönekonto), vacation account (semesterkonto) and home owner’s account (bokonto). In the grammar were also investment account and miscellaneous account (fondkonto and allkonto, respectively), and the savings account was often mistaken for the miscellaneous account. See the confusion matrix in Figure 8.

- Phonetically close words with different
meanings should be avoided. The most apparent example in the CTT-bank study was that an English “no” was added as a synonym of the Swedish “nej”, which led to some cases where the concept “yes” was recognised as “no”. See the confusion matrix in Figure 7. But the opposite (“no” interpreted as “yes”) was never the case, perhaps because more care had been taken in the choice of synonyms of yes to avoid this kind of error. Accordingly, no transactions were erroneously made because of recognition errors in connection with confirmation questions.

- Nearly identical expressions should not be used with different meanings in different steps of the dialogue. In CTT-bank, the expression “lista konto” sometimes means “give a list of my accounts” and sometimes “give a list of recent transactions for one of my accounts”, which might have confused users.

Table 6 presents concept error rates for each of the utterance types with answers to questions. Results are shown for three groups of utterances: for lexical and non-lexical utterances separately, and for the pool of lexical and non-lexical utterances. Within each of the three groups, error rates are given both for the first attempt test set and for the all attempts test set. Results for lexical utterances give information about the performance of the speech recogniser, while the pooled results show how well user input was understood by the demonstration system as a whole. Results for non-lexical utterances indicate how the system reacts to ill-formed input (with respect to the grammar authors’ expectations).

The overall concept error rate for lexical utterances is 12.0% for the first attempt test set and 13.1% for the all attempts test set. The corresponding results for non-lexical utterances are 62.3% and 61.2% respectively, but these utterances range from very close to lexical (for example “for the vacation account” instead of “from the vacation account”) to totally OOV/OOG utterances (for example “I said I don’t want to hear one more transaction”). Error rates for non-lexical utterances are of course very high. But error rates less than 100 % in this case means the application understood more than its speech recognition component. The shortages in speech recognition grammars are thus partly repaired by the application’s ability to handle some of the non-lexical utterances. As the non-lexical utterances are few compared to the lexical utterances, bad recognition results do not influence the total error rates that much. The concept error rate for all utterances, lexical or not, are for first attempts 20.6% and for all attempts 20.8%.

Recognition results on a per-speaker basis are presented in Figure 10, where they are also compared to speakers’ opinions about speech recognition performance (see Section 5.4.1).

### 5.2. Speaker verification

This section presents the results from the initial evaluation of the speaker verification component of the system. We present results that are a direct outcome of the user trial, including observations on false rejections and on the
number of attempts required for a user to get accepted by the system. It does not include observations on the false acceptance of impostor speakers, simply because no impostor attempts were made. A forthcoming publication (Melin, to appear) will present additional results and analyses, including results from simulated impostor attempts and backing results on another database.

In the verification stage of login, a speaker verifier and a speech recogniser were used to implement a combined text and voice verifier. A spoken utterance passed text verification if the prompted password string was included in any of the up to ten hypotheses produced by the speech recogniser, and it passed voice verification if the utterance was accepted by the speaker verification system. For clients to pass the entire verification stage of login, they had to produce an utterance that passed both text and voice verification within three attempts.

In 64 out of 72 calls where a subject tried to access the bank service the subject got as far as the verification stage of the login. In 3 of the 64, the subject was denied access after the third attempt. Since the subjects claimed to be themselves in all cases (no imposture attempts), this corresponds to a false reject rate of 4.7% for the combined text and voice verification process at the call level. On average 1.33 verification attempts were used per call (85 attempts in total). Table 7 shows false reject rates at one to three maximum number of attempts. The table shows error rates for the combined verifier as well as for the separate text and voice verifiers. Note that the results for the text verifier are fully independent of the operation of the voice verifier. The results from the voice verifier, on the other hand, depend upon the output of the speech recogniser used in text verification, as described in section 2.5.4.

In three of all the 85 attempts, the password was incorrectly spoken by the subject. In two cases the utterance was correctly rejected by the text verifier (and hence the combined verifier). The third utterance was falsely accepted by the text verifier and falsely rejected by the voice verifier (since the voice was in fact the correct one), and thus was rejected by the combined verifier. The combined decision was deemed correct because the password was incorrectly spoken. This explains why the false reject rate is lower for the combined verifier than for the voice verifier in Table 7 for max two attempts.

5.3. Dialogue

Figure 9 shows the division of time spent in each of the four sub-dialogues in an enrolment call: user authentication, getting data for speaker verification enrolment, identification number selection, and the choice to hear a demonstration call or not. One-way information from the system to the user is shown separately in the figure. This includes the welcome message and instructions in the beginning of the call, the goodbye message at the end, and informational prompts within the call that are presented before the first question in each sub-dialogue. The total duration of the 21 successful enrolment calls, excluding the optional demonstration call,
varied from 3:04 to 5:21 minutes, with an average of 3:55.

The login dialogue in bank service calls consisted of three parts: a greeting, “Welcome to CTT-bank”, followed by the identification and verification sub-dialogues. The identification sub-dialogue nominally consisted of the single short prompt “Say your name and your code” and a user response. If additional attempts were required, the system generated longer prompts with the aim to guide the user to say the correct phrase. This is in line with the goal to streamline dialogues for the experienced user (Ihse, 2000). In each call, the user was allowed a maximum of five attempts at identification. Correspondingly, the verification sub-dialogue nominally consisted of the prompt “Say the following digits: 9 6 8 9”, where the four digits were randomised for each attempt. If a second attempt was required (max three were allowed), the prompt was “The system did not recognise your voice. Say the following digits instead: 1 2 3 4”. Hence, in both the identification and verification sub-dialogues, additional attempts took longer time than initial attempts.

Among calls with successful logins, the duration of the identification sub-dialogue varied between 6 and 11 seconds where a single attempt was sufficient (average 8 seconds). Each additional attempt added between 17 and 24 seconds to the duration (average 20 seconds). The duration of the verification sub-dialogue was between 8 and 15 seconds for a single attempt (average 10 seconds), with an extra 11-21 seconds for each additional attempt (average 13 seconds). The total duration of the login dialogue, including the three second greeting, varied from 0:17 to 1:31 minutes, with an average of 0:33.

5.4. User aspects

Several user aspects were discussed in Ihse (2000), such as the time required to perform the various banking services, user opinions about several aspects of the system and an analysis of error recovery from the user’s point of view. This section presents additional results related to the use of speaker verification. In particular, it presents comparisons between the time required for enrolment with user opinions on enrolment, and between objective and subjective speech recognition error rates.

5.4.1. Speech recognition

Figure 10 shows speech recognition error rates per subject compared to each subject’s rating of recognition accuracy on a scale from 1 to 5, where 1 is worst and 5 is best. The rating shown in the figure is the mean value of answers to
follow-up questions number 6 and 16, “In your opinion, how well did the computer understand what you said during the registration?” and “In your opinion, how well did the computer understand what you wanted to do when in the banking service?” The complete questionnaire and most of the answers are included in Ihse (2000). Recognition error rates presented in the figure are weighted averages of all recognition results (digits, names and answers to questions) on each of the first attempt test set and the all attempts test set. The error rate presented in the figure for a certain speaker is the total number of recognition errors committed for that speaker, divided by the total number of recorded files for the same speaker. We use a heuristic definition of the total number of recognition errors based on word error rates on the one hand, and concept error rates on the other: the total number of recognition errors is the number of word errors in name and digit utterances plus the number of concept errors in the remaining files. The correlation between recognition error rates and subjects’ grading of them is apparent but not overwhelming.

5.4.2. Enrolment and login

Table 8 summarises the answers to follow-up questions about the enrolment call and the login phase of calls to the bank service. The high ratings on the login-related questions suggest that subjects were quite positive to the design of the login phase. The use of a customer’s own name in combination with a user-selected number seems to be well appreciated, and subjects found it easy to repeat the four-digit prompted password. The latter result is consistent with the finding in Lindberg & Melin (1997) that five digits are often difficult to repeat, while four poses no problems.

Ratings to questions about registration in Table 8 are slightly lower than for questions about login, but they still indicate that the design of the enrolment call was acceptable. The need for possibly lengthy enrolment is often presented as an argument against the use of speaker verification, but it seems that our subjects found the time spent on enrolment acceptable. Out of the average 3:55 minute enrolment call, 2:00 was the average duration of the sub-dialogue for eliciting the approximately 30 seconds of speech used for training a speaker model. The duration of the sub-dialogue varied between 1:02 and 3:05 minutes for the 21 subjects who succeeded with enrolment. Figure 11 plots answers to the two enrolment questions against the total duration of the enrolment call for each of the subjects (that correlates well with the duration of the speaker verification (ASV) data input sub-dialogue). As expected, Figure 11a shows a dependence between subjects’ opinion about the enrolment length and the actual duration of it. However, from Figure 11b the duration seems to have little influence on the subjects’ overall impression about the enrolment call.

The main source of variation in the duration of the enrolment call was re-prompts. To fill the ten slots in the ASV data input sub-dialogue, between 10 and 17 prompts were needed, with

<table>
<thead>
<tr>
<th>question</th>
<th>1 meaning</th>
<th>5 meaning</th>
<th>average</th>
<th>stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q9) What is your overall impression of the registration?</td>
<td>very bad</td>
<td>very good</td>
<td>3.76</td>
<td>0.77</td>
</tr>
<tr>
<td>(Q7) What did you think about the time taken by registration?</td>
<td>it was far too long</td>
<td>it might as well have been longer</td>
<td>3.14</td>
<td>0.79</td>
</tr>
<tr>
<td>(Q14) What is your overall impression of the login?</td>
<td>very bad</td>
<td>very good</td>
<td>3.90</td>
<td>0.70</td>
</tr>
<tr>
<td>(Q11) What did you think about the login system in which you stated your name followed by a personal code?</td>
<td>–”–</td>
<td>–”–</td>
<td>4.19</td>
<td>1.03</td>
</tr>
<tr>
<td>(Q12) How difficult was it to remember the four digits you were to repeat during the login?</td>
<td>very difficult</td>
<td>not difficult at all</td>
<td>4.57</td>
<td>0.81</td>
</tr>
<tr>
<td>(Q13) If you compare voice recognition as a method of login to PIN codes, which is the better solution?</td>
<td>PIN codes</td>
<td>voice recognition</td>
<td>4.00</td>
<td>1.05</td>
</tr>
</tbody>
</table>
an average of 11.5. An extra prompt was generated if the expected digits were not within any of the up to ten hypotheses from the speech recogniser. Other sources of variation were the speech detector failing to correctly detect utterance end-points, varying speech recognition processing times, and varying speaking rate among subjects. The average time per prompt-response cycle in the ASV data input sub-dialogue, including the time for speech recognition, ranged from 6.2 to 12.7 seconds among subjects, with an average of 10.3 seconds. Note that extra prompts, those asking for repetitions of an utterance, were longer than the initial prompts on each slot.

6. Previous work

Similar studies with speaker verification and telephone banking have been done within the CAVE, PICASSO and SPOTLIGHT projects as was noted in the introduction. They were the CAVE trial with Swiss Ubilab/UBS (CAVE consortium, 1998), the PICASSO Nationwide trial (Coutts, 2000), the PICASSO Fortis trial (Gu et al., 2000) and the SPOTLIGHT evaluation conducted at the University of Edinburgh. This section reviews the four studies.

6.1. Arrangement

CAVE/Ubilab: Subjects called a fully automated telephone banking service where they could check account balance, list recent transactions and transfer money between their own accounts. Speaker verification was used for user authentication in combination with a PIN. Bank accounts and money were fictitious.

PICASSO/Fortis: Subjects took part in a speech-driven investment game by calling an automated game server. Access to the server was protected by speaker verification. The task of the participants was to maximise the value of their portfolio by selling and buying stock. A helpdesk was set up to support the trial.

PICASSO/Nationwide: Speaker verification was used as a preface to the bank’s normal call-centre-based telephone banking service, but users still had to use their normal password to gain access to any service. Customers were never denied access based on the result of speaker verification, nor were they informed if they had passed or failed the verification test. Impostor testing was carried out after the main trial was closed.

SPOTLIGHT: The experiment was divided into data collection and verification testing. During data collection subjects called automatic recording dialogues that asked the caller to speak all the words and phrases listed on a personal details sheet, plus the words “balance” and “savings”. The personal details sheet contained information that would normally be available in (a British) telephone banking application. (It seems that in the UK, customer authentication in most current telephone banking services are based on the caller saying secret information, such as a certain part of the PIN or a password, an account or membership number, or the date of the most recent transaction.) A speech recogniser checked that the caller said the requested words and re-prompted when necessary, but no online speaker verification
was used. Using the resulting speech database, extensive speaker verification testing was later performed offline.

6.2. Enrolment

CAVE/Ubilab: Enrolment was made by two telephone calls.

PICASSO/Fortis: The first five valid calls to the service were used for enrolment. In each call, the user spoke a 7-digit account number that was used for training the speaker model. Hence, approximately 20 seconds of speech spread over five calls were used for enrolment. Calling line identification (CLI) or a PIN was used to authenticate the user during enrolment. CLI was used if the user called from his own mobile phone, otherwise a PIN was used. The same dialogue was used during enrolment and access calls. Hence, this is an example of “soft” enrolment, where customers use the service from the first call and are enrolled incrementally. 89% of the 109 participants who tried at least one enrolment call completed all five.

PICASSO/Nationwide: A single enrolment call was used. Each customer was contacted by an Enrolment Manager who verified the customer identity, explained the service and the enrolment procedure. The enrolment procedure itself was fully automatic. A speech recogniser was used to check the validity of the spoken utterances. On average, each user spoke 5 groups of 5 digits 5 times, providing approximately one minute of speech. The exact number of repetitions varied between calls depending on background noise and other factors. 94% of the 428 customers who tried to enrol succeeded in one or two attempts, while 6% failed to enrol. For 87% of the customers, enrolment took less than five minutes, while for 4% it took more than 10 minutes.

SPOTLIGHT: During data collection, 5 repetitions of each of the words and phrases on the personal details sheet, and the words “balance” and “savings”, were collected in a single dedicated enrolment call. The sheet contained 3 digit utterances (a 9-digit membership number, an 8-digit account number and a 4-digit PIN) and 3 words (a surname, a code name and a bank security word). The total average duration of the 3 digit utterances was 7.4 s and that of the 5 words 3.1 s. During verifier testing 1 to 3 repetitions of various combinations of the words and phrases were used. The average duration of the longest enrolment set used in testing (3 repetitions of all phrases and words) was approximately 30 s.

6.3. Login design

CAVE/Ubilab: Users spoke a 9-digit account number that was used for both identification and verification. In addition to the account number, they also spoke a 5-digit PIN.

PICASSO/Fortis: A spoken 7-digit account number was used for both identification and verification. Calling line identification (from the customer’s own mobile phone) or a 5-digit PIN (from telephones in the fixed network) was used as an additional authentication measure. The customer was given three login attempts per call.

PICASSO/Nationwide: Customers first had to enter their account number for identification, either by DTMF or by voice, and then speak a 5-digit PIN for verification. Speaker verification was performed on the PIN only. For the account number, 78% of callers chose to use DTMF and only 22% chose to use speech.

SPOTLIGHT: During data collection, subjects called up to four times to the service number and recorded one example of each of the digit phrases and words on the personal details sheet and the words “balance” and “savings”. There were at least four days between each call. During offline verifier testing, various combinations of the above words were used. The average total duration of the longest test set (a combination of all the words and phrases) was 10 seconds. For administrative purposes, a 5-digit reference number was used to identify the caller during data collection. The issue of identification was not addressed in the study.

6.4. Size of trial

CAVE/Ubilab: 185 speakers enrolled into the system and 4363 access calls were made. In 73% of the calls speakers tried to access their own account, while in the remaining 27% they tried to access another user’s account.

PICASSO/Fortis: 109 subjects called the service at least once. 98 of them completed the five enrolment calls. The field test was open for five weeks, during which a total of 2662 calls were made to the system (including enrolment calls). In addition, 1230 impostor attempts were made against 50 accounts for which there were also many calls made by the true user.

PICASSO/Nationwide: The pilot system was open for 7 months during which 428 subjects attempted to enrol. 827 calls were made where the caller started a verification test. In addition, 423 impostor attempts were collected after the trial.

SPOTLIGHT: 1514 speakers were recruited for data collection and 995 completed all five
calls. Some calls were discarded as incomplete after the data collection phase and therefore a subset of 779 speakers were used for the main series of verification tests.

6.5. Realism
CAVE/Ubilab: Users were assigned fictitious bank accounts and could perform the same services as in CTT-bank. A game was arranged to motivate users to call often. Each day some amount of money was randomly placed in one of the user’s savings accounts and was left there for 24 hours. The user could keep the (fictitious) money by transferring it to his private account, where money would accumulate. At the end of the trial, users were given a number of lottery tickets proportional to the balance on their private account. Real money was awarded to winners in the lottery. All participants who at least enrolled into the system also received a pre-paid telephone card. Participants were mainly bank staff members.

PICASSO/Fortis: Subjects were employees of a financial institution. Their motivation was the investment game where no real money was involved.

PICASSO/Nationwide: Data was collected in calls to a real banking service by regular customers. However, customers had to use their normal password to gain access. They were aware that the result of speaker verification was logged, but they never knew what the result was. To motivate subjects, they were given a toll-free telephone number throughout the trial and would take part in a prize draw at the end of the trial if they called the service at least five times.

SPOTLIGHT: The conditions for the study was more like most traditional research experiments in speaker verification in the sense that testing was performed on data recorded with no online feedback from a speaker verification system. However, the recorded material is relevant for banking applications, and the recording dialogue was designed to resemble a real banking dialogue.

6.6. Results of the trial
CAVE/Ubilab: An unexpectedly large percentage of users experienced a high rate of false rejections. Technical and/or operational problems with enrolment were suspected as the reason, and this issue was later addressed within the PICASSO project. No numbers on speaker verification error rates are disclosed in the public report, but it is stated that results were disappointing and differed much from results from earlier research experiments within the project. A summary of user reactions based on a questionnaire given to all users after the trial states that the enrolment procedure posed “hardly any problems” for users, and that time spent on voice registration was judged as reasonable. The length of the access procedure, however, was judged as too long.

PICASSO/Fortis: The false reject rate for the speaker verification system during the trial was 2% at the call level (rejects after up to three attempts), where all false rejects appeared for 19% of the subjects. A false accept rate of 0.6% was observed. The subjective evaluation made by subjects showed that they were very pleased with the service, and that they preferred a combination of a PIN and speaker verification to speaker verification alone. 92% of the subjects would like a human operator to backup the automatic system in case of problems.

PICASSO/Nationwide: Using the collected impostor trials, an EER of around 12% was calculated initially. With improvements made to the speaker verification system after the trial, EER on recorded data was reduced to 3.1%. Customers found the enrolment procedure tedious, and the report suggests that incremental enrolment over many calls should be used instead. It was also noted that a fallback strategy must be considered for handling failed enrolments. Verification error rates were found to be too high, but it was noted that speaker verification could be used as a confidence indicator and as an additional means of security in the context of a call-centre based application. Concern was also expressed about the use of a fixed PIN that can potentially be overheard and even recorded.

SPOTLIGHT: Half Total Error Rates (HTER, the average of false rejection and false accept rates for a given threshold setting and in some sense comparable to EER) were presented for an extensive set of conditions. For example, using the longest enrolment data set (around 30 s), HTER with same-sex impostors was 1.4% when a verification test was made on a combination of the account number, the membership number and two digits from the PIN. The corresponding error rate was 3.4% if the five words were used instead, and 1.2% if the digit phrases and words were combined. In addition to the main series of tests with 779 speakers, separate test series were made with related speakers and with deliberate mimicry. It was shown that the false accept rate was much higher (around 50% with the decision threshold used above) for an identical twin impostor than for a random, unrelated impostor of the same gender. False accept rates for parent/child was
also higher, while for non-identical siblings the difference was slightly less. The mimicry study indicated that impostors who were allowed to listen to their target speaker before attempting to mimic him (with no feedback on how good their imitation was) were able to increase their chance of being (falsely) accepted. Similar results on mimicry with the verifier used in CTT-bank were found in Elenius (2001), where it was also found that training with feedback from the verifier was an additional help for the impostor.

6.7. Other notes
In the CAVE/Ubilab trial, each user was given account numbers and modified PINs for five other users of the same gender. The modified PIN was used to detect if an access attempt was made by the true user or not.

7. Discussion

7.1. Speech technology for telephone banking
The first of the primary questions put in this study was “Is speech technology useful for telephone banking?”. Our understanding is that the answer is yes. The test subjects were pleased about the services, despite the technical problems. They had hardly any problems in understanding the speech from the synthesiser, and the voice login procedures were much appreciated. With a faster system, allowing users to speak their intentions in one single sentence (for example “transfer 250 SEK from my salary account to my savings account” instead of “transfer”, “salary account”, “250 SEK”, “savings account” with prompts in between), the possibilities that speech technology offer, would probably be even more appreciated.

The service portfolio we have chosen – transferring money and information about accounts – is apparently working well also without speech synthesis, speech recognition and speaker verification, as most banks in Sweden have DTMF-controlled Interactive Voice Response (IVR) systems for these services today. What the mentioned speech technology components can add to these systems are among other things:

- easier access in situations when pushing buttons on a telephone device is impractical
- easier access for people having difficulties pushing buttons, for example the visually or motorically impaired
- less information to remember, for example PINs and account numbers
- speed, if the dialogue allows filling many slots at the same time
- improved security compared to if PINs are used and an impostor knows the PIN. Improved over-all security if speaker verification is combined with a PIN.

7.2. Dialogue design
The second of the primary questions put in our study was: “How should a dialogue be designed to work well for users?”. After examining spoken input in CTT-bank and subjects’ questionnaire replies, we give the following guidelines on this:

- Consistent: an utterance or command should mean the same thing everywhere in the dialogue flow
- Intuitive: the natural ways people use to express their intention should work
- Flat hierarchies: one utterance should suffice for one service
- Effective and fast: long calls cost money for both the bank and its customers
- Flexible: if customers change their minds or something goes wrong, the system should handle that in a nice way
- Enough but not too many confirmation prompts: the confirmation of what was said must not disturb the flow of the dialogue, but the customer must be confident that the system has correctly understood what was said
- Good balance between stability and adaptability: users should feel familiar with the application every time they use it, but the experienced user should not be bothered with unnecessary information that only suits beginners.

7.3. More design issues for research and commercial use
Our service portfolio and dialogue design are well suited for running DTMF and speech input in parallel, adding a modality to today’s telephone banks. However, there is a trade-off between such parallelity and a dialogue design well suited for speech technology. A system designed for use without DTMF would be

---

12 Except in the verifier in Elenius (2001) gender-dependent cohort models were chosen based on similarity with enrolment data rather than test data.
designed differently than a system with both DTMF and speech. With DTMF and spoken input in parallel, the possibilities of speech technology are severely limited and can not be completely made use of. For example, speech technology provides the possibility to offer many alternatives (long menus) at a single point in the dialogue, and can allow the user to provide complex information in a single utterance. None of these are practical in a dialogue with DTMF input.

Our main reason to refrain from using DTMF in CTT-bank was our explicit wish to investigate “pure” speech technology. For commercial use, the considerations would be different, as there are also advantages in using DTMF and speech in the same system. DTMF backup is good when calling from noisy environments that complicate speech recognition. Systems that combine DTMF and speech allow service providers to offer their customers one single application with a perhaps already known telephone number and facilitate for users who want to try new technology without extra effort.

Good guidance for decisions on using speech input/output are given in modality theory and theories on speech functionality (Bernsen et al., 1998) through systematic evaluation of claims on applications.

Probably, telephone banking application areas other than ours would benefit more from using speech technology. One example is stock trading, where names of new shares and funds can be easily added to the system with no need for new recordings or redesign of menu hierarchies. Stock trading requires a larger vocabulary than the services in CTT-bank. The large vocabulary is another reason to choose speech technology over DTMF, but also requires more from the speech recogniser in terms of low error rates.

Another design/development aspect which is important both in research and in commercial use is logging of incoming calls. We believe that logging is equally important in research and commercial use, but the requirements on what information to collect are probably different.

7.4. Speech recognition

Many of our grammars were designed to recognise pre-defined keywords in pre-defined carrier phrases, with the goal to allow subjects to speak with a relatively free language. Given the rather simple, keyword-based language “understanding” component of the CTT-bank system, conventional keyword spotting would have been an obvious alternative. With keyword spotting, where the recogniser looks for pre-defined keywords in arbitrary carrier phrases, the same distinction between lexical and non-lexical utterances would not have appeared as with our grammar-based speech recognition. However, our choice of grammar-based speech recognition was made with a future, more complete language understanding in mind, where every word spoken by the user would be taken into account. For example, a proper language understanding component should be able to tell the difference between “from the savings account” and “to the savings account”, and even be able to pick two slots of information from the phrase “from the salary account to the savings account”. It should also be able to understand utterances like “no, not the savings account”. A keyword spotter may not be able to extract sufficiently detailed information about the spoken utterance to allow this.

When creating grammars and vocabularies for grammar-based speech recognition, one must try to predict what users will say. This is usually done with data-driven methods and an iterative design process: example speech is collected from users; observed utterances are included in the grammar and the vocabulary; more examples are collected with the improved system; grammars are updated, etc. During this initial trial with CTT-bank, we have collected data for the first round of this iteration process.

Even with good grammar and vocabulary design, there will always appear utterances that cannot be recognised completely because of new words, false starts, hesitations, etc. A good confidence measure from the speech recogniser is then a key to robust recognition. The confidence measure associated with a recognition output string should indicate if the string is correct, or if some words, or the entire utterance, may not have been represented in the grammar. This would allow a dialogue system to fail gracefully instead of moving forward on incorrect information, as was sometimes the case in the CTT-bank trial, where a confidence measure was not available from the speech recogniser.

7.5. The use of speaker verification

7.5.1. Enrolment

The need for possibly lengthy enrolment has often been seen as a disadvantage of speaker verification compared to PINs, and was one of the major concerns in the beginning of the CTT-bank project. In the user trial, 30 seconds of speech from a single call was used for building speaker models. It took around two minutes to
collect this speech, as part of a four-minute registration call. In the subjective evaluation, our subjects did not find this time too long. A similar reaction was found in the CAVE/Ubilab study, where subjects found the time spent on enrolment reasonable. In the PICASSO/-Nationwide trial, however, subjects found the enrolment tedious. The duration of the majority of their enrolment calls was comparable to the length of the entire registration calls used in CTT-bank (about half of which was collecting samples for speaker verification enrolment). A notable difference between the Nationwide trial on one hand, and CTT-bank and the other trials on the other, is that Nationwide used real bank customers as subjects while the others used mainly employees of banks and a technology supplier. It is quite possible that bank customers in general are more critical to spending time on enrolment. It is also quite possible that particular groups of customers are more motivated than others. For example, visually or motorically impaired people may find speaker verification particularly useful compared to typing PINs, and therefore are willing to spend more time on enrolment.

A suggestion from the Nationwide trial was to use some sort of incremental enrolment instead of a single long enrolment call, such that enrolment data is collected bit by bit from ordinary bank service calls. This was also tried in the PICASSO/Fortis trial, where there was no enrolment call in the traditional sense. With incremental enrolment, some alternative means of authentication must be used until enough enrolment data has been collected. PINs and calling line identification are two possibilities that were also used in the Fortis trial. It seems that subjects in the Fortis trial were positive to incremental enrolment.

7.5.2. Performance

The error rate in speaker verification depends upon the amount of speech available for enrolment and test. This has been confirmed by the Spotlight consortium (2000) and several other studies. In the CTT-bank study, too few subjects and sessions were recorded to allow conclusive remarks on the absolute level of error rate for the used verification system. In a forthcoming paper (Melin, to appear), the results will be further analysed with the backing of a larger database of recordings (Gandalf) and will include results on simulated false accept rates and EERs. In the present paper, a false reject rate of around 5% has been reported. This error rate was observed with around 30 seconds (10 times 5 digits) of enrolment speech, a 2-second test utterance (four prompted digits) and a maximum of three attempts per call. The false reject rate at one attempt per call was as high as 20%. The EER for the same verifier tested on the Gandalf database (Melin, 1996) with same-sex impostors and equivalent enrolment and test conditions as in CTT-bank is 8.2%. This suggests that the false accept rate in CTT-bank, if measured, would have been smaller than the false reject rate. Comparing to the Spotlight study, a test case with a similar amount of enrolment data (3 times 19 digits) can be found. HTER for the Nuance Verifier with same-sex impostors was then 2.2% with verification on an 8-digit account number and 5.9% on 2 PIN digits (Spotlight consortium, 2000; Table 15, condition amp_3). The corresponding error rate for 4 digits, the same number used in the prompted password in CTT-bank, can then be interpolated to approximately 4%. This number corresponds to a maximum of one attempt, though no actual verification was used during data collection in the Spotlight trial (as was the case in the collection of the Gandalf database). Slightly older versions of the Nuance Verifier and the verifier used in CTT-bank have previously been compared directly on the Polycost database, where they produced similar error rates on a digit-based task (Nordström et al., 1998).

Results in the previous paragraph indicate that the EER/HTER in a real application would be in the vicinity of 5% for the amounts of enrolment and test data used in the CTT-bank trial. This is clearly too high for most banking applications. Possible ways to reduce the error rate are:

- Use more data for verification testing. All words spoken during a service call could be used, such as account names, amounts and command words. Since the length of the test utterance in the CTT-bank trial was very short, increasing it is likely to yield a considerable reduction in error rates. Additional enrolment data for new words and phrases could be collected incrementally, or the enrolment call might have to include other utterances than digit sequences.
- Collecting better training data. Online supervision of data collection during enrolment may improve speaker model quality and balance the uneven distribution of errors among customers often seen in speaker verification (Gu et al., 2000).
- Collecting more training data. This probably has to be done incrementally during service
calls since prolonging the enrolment call may reduce customer satisfaction.

- Improving the verification system. The field is actively researched and commercial verifiers improve continuously.

HTERs around 1% were for instance found in the Spotlight trial with the Nuance Verifier using more enrolment and test data than in the CTT-bank trial, showing that such low error rates are achievable.

Another possibility for applications with strong demands for high security and low customer “insult rate” is to combine speaker verification with other means of authentication, such as PINs and calling line identification. In this case, the decision threshold would be adjusted to aim for a very low false reject rate at the cost of a higher false accept rate. The total false accept rate would still be low because of the combination with a PIN. The main function of speaker verification is then to improve security when an impostor knows a customer’s PIN.

8. Conclusions

The main conclusion from the CTT-bank study is that speech technology is useful for telephone banking. Although the CTT-bank application was far from perfect, subjects managed to use the available bank services and were positive to the application. This indicates that our dialogue design was useful, though not necessarily optimal. In the discussion, we have suggested design criteria for improved dialogue handling. We feel, however, that dialogue design and dialogue management need more attention than given so far in this project.

As for speech recognition, it seems that even not-so-good recognition can work well in a dialogue (as opposed to for example in a dictation situation), if only the error handling is good. Extra knowledge about what spoken input to expect, for example restrictions on valid ID numbers at enrolment, helps choosing the most likely alternative from an n-best list. A certain amount of explicit confirmation should also be included in the dialogue. Our relatively high speech recognition error rates can partly be explained by our use of domain-neutral acoustic models trained on read speech. Domain-dependent acoustic models would perform better, and the additional cost for creating them would be motivated in a real application.

The length of the enrolment call required by the use of speaker verification for customer authentication appeared less critical to our subjects than anticipated before the trial. However, comparisons with other similar trials suggest that this may be a result of our selection of subjects. Bank customers in general may be more critical to spending time on enrolment.

We tried speaker verification with a very short, prompted test utterance. It was accepted by subjects from a usability point-of-view, but the observed error rates were too high for a banking application. With more training data collected through incremental enrolment, and more test data through the use of additional utterances that appear naturally during the dialogue, we argue that equal error rates in the order of 1% are achievable for speaker verification alone. We also pointed out the possibility of combining speaker verification with other means for authentication, such as a PIN. The use of the customer’s name together with a number for claiming an identity also seems to have been successful from a usability perspective.

Based on our experiences, we are confident that speech technology is a key to more capable and user-friendly telephone banking services in the near future.

9. Acknowledgements

This research was carried out at the Centre for Speech Technology (CTT), a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

10. References


