O-TEL - An Experimental Reverse Directory Telephone Service with Barge-In Capability

Andrej Miksic, Zdravko Kacic, Bogomir Horvat
Faculty of Electrical Engineering and Computer Science
University of Maribor
Smetanova 17
2000 Maribor, Slovenia
andrej.miksic@uni-mb.si

ABSTRACT

Issues of practical implementation of a telephone dialog system are discussed with the reference to the experimental O-TEL system (a telephone reverse directory service) developed at University of Maribor.

The system offers a user to call the system, utters the telephone number in order to receive information on the subscriber (name and address).

Critical implementation issues are: allowing a full-duplex telephone interaction (user barge-in); design of a dialog flowchart that enables fluent and fast dialog; robust feature extraction with channel compensation; assuring a high recognition accuracy; and synchronizing the system prompt signal with the feature extraction part of a recognizer for the purposes of introducing echo cancellation as a part of feature extraction.

Field trials have shown especially high rate of dialog completion and a good error recovery scheme. The capability of barge-in reduced the dialog completion time for up to 56\% (especially in error recovering phase).

I. INTRODUCTION

Development of automatic spoken dialog systems is now heading towards fully deployed applications operating in real-time. Systems operating in laboratory environment are often lacking the criteria, that are important for practical implementation. Such systems only aim to reach the highest possible recognition rates for speech databases used in testing. In this way, communication efficiency and dialog effectiveness too often play the inferior role. This is the reason why the disparity between laboratory systems and fully deployed applications is still relatively large.

Efficient speech preprocessing, including channel compensation and echo cancellation for assuring a full-duplex communication, represent a necessary condition for the effectiveness of a speech dialog system, which is important for overall dialog usability.

In this contribution we deal with the possibilities of employing the echo cancellation for achieving a full-duplex communication, and with the advantages that such a dialog mode can offer.

II. A BARGE-IN PROBLEM

One of the primary goals of the O-TEL system was to exhibit an option of simultaneous speech input and output (referred to as “barge-in”, “talk-trough” or “user priority mode”) [5].

Because of echoes generated in hybrid circuits in the telephone terminals and central offices (see Fig. 1), a sampled received speech that is an input to a feature extraction part of a recognizer contains not only a speech signal from the user, but also an echo signal of a system prompt, thus greatly degrading the recognition performance and usability of the dialog system.

The problem is most difficult to handle during a barge-in, when both user and system prompt are active at the same time.

To be used in a real-time system, the subband echo canceller was utilized in a way not to add a significant computational overhead to a speech recognizer and to avoid degradation in recognition performance of a speech recognizer. The frequency subband signals were led directly to the feature extraction part of a recognizer.
There are different possibilities of system reaction during a barge-in. It may be deaf during the output, not reacting to input at all; or it can react by stopping the system prompt and respond according to user barge-in. Input speech sampling in the O-TEL system is active all the time after the first system prompt has been started and therefore allows a user barge-in.

A demand to stop the system prompt may be triggered by the double-talk detection or by the speech recognition (when something has been recognized).

III. DIALOG DESIGN

O-TEL is a reverse directory assistance telephone service. The system offers a user to speak a telephone number in order to receive information on the subscriber.

The O-TEL dialog was designed with a constraint to meet the following requirements:

- Allow a user to barge-in and react according to speech recognition.
- Fast and accurate information query.
- Error recovery such as to always lead to a dialog completion.
- Clear and non-repeating system prompts.
- Robustness to environmental and line noises.

Fig. 2 shows a dialog flowchart of the system. System messages are composed using a word-based speech synthesis that also account to some prosodic timing rules.

System messages on Fig. 2 are shown as ellipses. The dialog can react to all the possible dialog situations. Some of the common ones are:

- User is silent for a longer time period. The system should warn a user to continue.
- Number spoken is not in the telephone directory.
- System misrecognized a number.
- Recognizer rejected the user utterance.
- Number spoken is incomplete.

Here is an example of the O-TEL dialog completion:

No user barge-in:

**System:** Hello, this is the automatic reverse-directory system O-TEL. Say the telephone number digit by digit and I will repeat each digit after your. If I repeat incorrectly, say “delete”.

**user:** two
**system:** two
**user:** two
**system:** two
**user:** zero
**system:** zero

It is important to define the system behavior in the double-talk situations. Two schemes are possible: when double-talk is detected, the system prompt could immediately be stopped and the system would only “listen” to the user. In the case of false alarm, such scheme may confuse the user. We decided on a more complex solution: the system question stops only when the system has recognized a word or a sentence – only then it stops and continues according to user response.

**user:** (waits a long while)
**system:** So far, I have heard the digits 220. Please continue.
**user:** (waits a long while)
**system:** The number 220 is incomplete. Please continue.
**user:** seven
**system:** seven
**user:** seven
**system:** nine
**user:** delete
**system:** I'm sorry. Please repeat the last digit. So far, I have heard the digits 2207.
**user:** seven
**system:** seven
**user:** nine
**system:** nine
**user:** (is quiet - therefore the last digit is confirmed)
**system:** The number 220-779 belongs to ...(name, address). Thank you for calling.

With user barge-in: (sign # indicates that the system prompt was interrupted by user barge-in)

**system:** Hello, this is the automatic reverse-directory system O-TEL. Say the telephone number digit by digit and I will repeat each digit after your. If I repeat incorrectly, say “delete”.

**user:** two
**system:** two
**user:** two
**system:** two
**user:** zero
**system:** ze#
**user:** seven
**system:** seven
**user:** seven
**system:** nine
**user:** delete
**system:** I'm sorry. Please repeat#
**user:** seven
**system:** seven
**user:** nine
**system:** nine
**system:** The number 220-779 belongs to ...(name, address). Thank you for calling.
IV. ECHO CANCELLATION – HANDLING THE DOUBLE-TALK CONDITIONS

The O-TEL system uses a subband echo cancellation scheme. A normalized LMS adaptive algorithm was used to generate a replica of the system echo and subtract the subband echo estimation signals. Subband division was carried out according to the mel scale. In the feature extraction part of a speech recognition system, the speech spectra is calculated. This gives motivation to perform echo cancellation on subband signals. Subband signals of estimated echo are subtracted from the subband signals of input speech (more details in [8]).

Sub-band adaptive filtering is also one of the solutions to increased computation and slow convergence associated with the conventional full-band approach [2]. Reduced computational load is due to a time decimation in sub-bands (time/frequency transformation is performed on signal time-frames, not after each sample). The system coefficients adapt and in the ideal case the error signal converges towards zero (adaptation is possible when only echo is present, without user speech). Step size factor of the NLMS algorithm is used to control the adaptation speed. The barge-in (double-talk) situation has to be detected in order to prevent miss-adjustment of the echo canceller (lower the stepsize). We developed a double-talk detection method by tracing the residual error function of the echo canceller [8].

The computational load of the subband echo canceller used in the O-TEL system represents a 45 % part of the feature extraction time.

V. PERFORMANCE EVALUATION

Prior to conducting the field trials, we simulated the recognition performance in double-talk conditions in order to tune the echo canceller and the double-talk detection parameters.

A. Simulations

The training and test databases were “Voice Mail” databases (digits and 6 commands) recorded over the German PSTN (analog and digital). Test and training sets are non-overlapping.

Results were obtained simulating the heavy double-talk situations by adding the echo signals to above
mentioned test database utterances. Results are summarized in the Table 1.

Recognition experiments were conducted using the baseline recognizer characterized below:
Isolated words were recognized using Continuous Density Hidden Markov Models with Laplacian density functions. For each frame (10 ms spaced) we extract 52 element feature vector. It consists of mel-scaled cepstral, delta cepstral, end energy components. We use context-dependent diphone models. Each phoneme consists of 3 segments.

<table>
<thead>
<tr>
<th></th>
<th>clean</th>
<th>+echo</th>
<th>+echo+ EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total WA</td>
<td>94.9 %</td>
<td>64.4 %</td>
<td>92.4 %</td>
</tr>
</tbody>
</table>

Table 1 – Recognition results

Recognition results for double-talk periods are greatly reduced due to the telephone line echo. In case of echo cancellation, the recognition is considerably improved.

B. Field trials

The real world pilot tests have shown that users used both intentional and unintentional barge-in. The decision to stop according to the recognizer status rather than on speech detection proved feasible.

Unexperienced users made little use of barge-in capability while a clear tendency was shown that experienced users wish to use the service as quickly as possible and have greatly exploited the barge-in capability, which reduced the average data query time for over 40%.

<table>
<thead>
<tr>
<th></th>
<th>No recog. errors</th>
<th>1 error recovery</th>
<th>2 error recoveries</th>
</tr>
</thead>
<tbody>
<tr>
<td>No barge-in</td>
<td>54 s</td>
<td>72 s</td>
<td>98 s</td>
</tr>
<tr>
<td>Barge-in</td>
<td>32 s</td>
<td>37 s</td>
<td>43 s</td>
</tr>
<tr>
<td>Time save</td>
<td>41 %</td>
<td>48 %</td>
<td>56 %</td>
</tr>
</tbody>
</table>

Table 2 – Average dialog completion times

Table 2 shows a time taken for the dialog completion in different dialog situations. Especially in error recovery mode, barge-in can greatly shorten the dialog completion time.

VI. CONCLUSIONS AND FUTURE WORK

Recognition results infer that it is necessary to include the echo canceller in a telephone dialog system that can deploy simultaneous speech input and output.

Barge-in proved to be an important issue in real-life voice driven telephone applications especially for experienced users. The system O-TEL performed well in pilot tests and will be offered to public use as an experimental system when ported to Windows NT environment connected to ISDN digital telephone network (planned for the second half of 1998). Then the extensive usability tests will be carried out.

A multi-lingual extension of the system allows a user to speak in any of the supported languages (Slovenian, German and English are currently supported). The system prompts adopt to the recognized language.

A work is currently been carried out to implement connected-digit, as well as naturally spoken number recognition. Apart from reaching a satisfactory recognition rates, the main implementation issues are here a new dialog design with different error recovery scheme and a robust echo cancellation scheme, effective for more prolonged double-talk periods.

REFERENCES