ON FIELD EXPERIMENTS OF CONTINUOUS DIGIT RECOGNITION OVER THE TELEPHONE NETWORK

D. Falavigna, R. Grettel
IRST — Istituto per la Ricerca Scientifica e Tecnologica
38050 Panté di Povo, Trento, Italy.
e-mail: falavi@irst.itc.it grettel@irst.itc.it

Abstract
In this paper a continuous digit recognizer over the telephone network in real time will be described. The activity has allowed the realization of a system, installed in some Italian telephone exchanges, for providing semi-automatic collect call services. Data collection has also been performed, and a field database was built. Either a continuous digit recognition task and a confirmation task, requiring rejection, have been defined. Recognition results are presented.

INTRODUCTION
The activity reported in this paper led to the realization of a system, installed in some Italian telephone exchanges. It provides two semi-automatic collect call services, called “Italy Direct” and “170”. These systems require the recognition of digit sequences, as well as of yes/no. In the last case rejection of unforeseen sentences must be used to assure sufficient robustness with respect to user inexperience.

To train and test the system some telephone speech databases, later described, have been used. In particular a database, called FIELD, acquired during the system usage, will be introduced and discussed.

Finally, results will be presented, both for the digit recognition task and for the confirmation task.

SYSTEM DESCRIPTION
“Italy Direct” and “170” services allow an Italian user (user A) to communicate with another Italian user (user B), without interacting with a foreign operator. User A is in a foreign country and user B is in Italy (service “Italy Direct”), or vice-versa (service “170”). Briefly, user A calls an Italian operator, introduces himself, and gives him the telephone number of user B. Then the operator calls user B and asks if he accepts a collect call from user A. If yes, user A and B are connected. The automatic system, called POA, can handle both phases, requiring the presence of the operator only when the speech recognizer fails. The POA (see figure 1) consists of the following basic modules:
- a supervisor which controls the interaction flow;
- an interface with the public exchange UT100;
- a speech recognizer for digit sequences and confirmations.

Figure 1: Services “Italy Direct” and “170”: UT100 is the public exchange, POA the automatic system. In case of failure a human operator intervenes.

A typical call is first passed to a POA, which interacts (possibly using DTMF) with user A. After some minor informations, the telephone number of user B is requested and digit recognition is performed. Then a confirmation is asked. If user A does not confirm the digit sequence after a maximum number of trials (usually 2 or 3), the call is passed to an operator, otherwise the POA calls user B, which has either to accept or to refuse the incoming call.

The speech recognition module is responsible for executing some microoutines, that basically allow to get some elementary information. Examples are GetUserNumber(), which has to prompt a message, get and recognize the user sentence, ask for confirmation, and possibly repeat the whole process for a programmable number of trials. It returns either a user-confirmed digit sequence or a FAIL. The user can speak only after the vocal prompt, i.e. there is no barge in capability. A start-end point detector with a dynamic threshold [2] is used to detect the speech signal, that can be also stored for further analysis. At present, a total of 104 POAs have been installed near Rome and Milan, and the whole system is under test by Telecom. System design and development, apart speech recognition, have been carried out by AT-System and Italtel.
SPEECH DATABASES
The speech databases involved in the training / evaluation of the system are the following:

- CLEAN, a band filtered version (between 300 Hz and 3700 Hz) of IRST databases, APASCI and
  SPK, acquired in an acoustically isolated room. APASCI [1] consists of 5,215 phonetically rich
  sentences (194 speakers). SPK was collected for
  speaker recognition purposes, and contains about
  30,000 digits (both isolated and sequences) uttered
  by 107 speakers. All this material is labeled and
  segmented in words and phonemes.

- PHONE, formed by 5,210 sentences (280 speakers)
  recorded through the public telephone network
  [9] by means of an automatic system that calls
  a previously advised speaker. Sentences include
  confirmations, digit sequences and phonetically
  rich sentences. This database has been manually
  checked and transcribed. Due to the inability of
  the speakers involved in the collection with speech
  recognition applications, several spontaneous
  speech phenomena (hesitations, breaths, false
  starts, etc.) are present in the recordings.
  Particular care has been taken for the transcription
  of these phenomena. Each file in the database
  has been classified into one of the following classes:
  - highQ, sentences without spontaneous speech
    phenomena;
  - medQ, sentences with some weak spontaneous
    speech phenomena (breaths, noises, hesitations,
    isolated laughs, etc.);
  - lowQ, sentences with strong spontaneous speech
    phenomena (false starts, speech and laughs
    together, etc.).

For what concerns confirmations, it has been observed that, even if the system explicitly required to
utter only "yes" or "no", a large percentage of
the answers differs from them. In particular, the database contains the following distribution of
yes/no answers:
- 64.1% clean yes/no;
- 19.2% yes/no with some weak spontaneous
  speech phenomena;
- 2.7% other expressions clearly meaning yes/no;
- 3.7% yes/no followed by other words (motiva-
  tions, comments, etc.);
- 10.3% expressions without a clear yes/no mean-
  ing.

- SIRVA, connected digit training material, provided
  by CSILT. It contains 4,372 digit sequences having
  length 2-4, uttered by 1,096 speakers covering all
  Italian regions.

- FIELD, speech material collected by the system in
  the public exchange during the service operation.
  These data have been divided according to their
  meaning (digits, yes/no of user A, yes/no of user
  B). At present, more than 6,000 files have been
  collected and checked, each one containing a single
  sentence; more than 2,000, mainly from users A,
  have been labeled as garbage. In the next section a
  description of these data will be given.

FIELD DATABASE
Data collected so far have been divided into three
groups, according to the month in which they were
recorded (December 1996, January and February
1997). The number of files collected in February is
larger than those collected in the previous months.
Each file has been manually checked, transcribed and
assigned to a predefined class, according to its
content. Several files have been classified as garbage, typically
those containing silence (the user either speaks
during the vocal prompt, or simply remains silent), or
those due to children that make calls just for fun (note
that the call by user A is free). In general, these data
can be considered hard to handle. Since many calls
come from public boxes, background voices and noises
are frequently present. Sometimes the user himself
comments to other persons what is going on.

<table>
<thead>
<tr>
<th></th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>digits</td>
<td>133</td>
<td>135</td>
<td>899</td>
<td>1167</td>
</tr>
<tr>
<td></td>
<td>(54.1%)</td>
<td>(65.9%)</td>
<td>(60.0%)</td>
<td>(59.8%)</td>
</tr>
<tr>
<td>tens</td>
<td>27</td>
<td>19</td>
<td>74</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>(11.0%)</td>
<td>(9.3%)</td>
<td>(4.9%)</td>
<td>(6.2%)</td>
</tr>
<tr>
<td>oov</td>
<td>21</td>
<td>5</td>
<td>88</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>(8.5%)</td>
<td>(2.4%)</td>
<td>(5.9%)</td>
<td>(5.8%)</td>
</tr>
<tr>
<td>garbage</td>
<td>46</td>
<td>46</td>
<td>428</td>
<td>549</td>
</tr>
<tr>
<td></td>
<td>(26.4%)</td>
<td>(22.4%)</td>
<td>(29.2%)</td>
<td>(28.2%)</td>
</tr>
<tr>
<td>tot</td>
<td>246</td>
<td>205</td>
<td>1499</td>
<td>1930</td>
</tr>
</tbody>
</table>

Table 1: FIELD: statistics on the collected digit se-
quences.

Apart from garbage data, digit sequences have been
divided into the following classes: digits (valid digit
sequences), tens, (sequences containing also tens,
instead of digits alone), oov (sequences containing
out-of-vocabulary words, false starts, etc.). Table 1 re-
ports some statistics. Note that only 60% of all the
files are valid digit sequences. We have estimated that
the children are responsible of about one fourth of the
garbage files. It is worth noting that the percentage of
sequences including tens is decreasing from about
10% (Dec, Jan) to 5% (Feb). This may be due to the
fact that users have learned to interact with the sys-
tem correctly. Out-of-vocabulary words include digit
sequences in different languages (English and Span-
ish), injuries ("0 7 6 5 7 or Cristo sto computer"),
comments to another person ("0 3 5 4 ora lo ripete
..." - "... now it repeat it ..."), explanations ("eh
devo telefonare in Francia ma non so il prezzo di Pa-
rigi" - "I have to make a call to France but I don't
know the prefix of Paris").

Tables 2 and 3 show statistics for the confirmation
task. Each file has been assigned to one of the follow-
ing four classes:

- yn: only "si" and "no", possibly with weak sponta-
nous speech phenomena;
- richyn: expressions clearly meaning yes/no ("si,
Table 2: FIELD: statistics on the confirmation task of user A.

<table>
<thead>
<tr>
<th></th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>yu</td>
<td>291</td>
<td>308</td>
<td>564</td>
<td>1163</td>
</tr>
<tr>
<td></td>
<td>(46.5%)</td>
<td>(47.1%)</td>
<td>(43.4%)</td>
<td>(45.1%)</td>
</tr>
<tr>
<td>richy</td>
<td>23</td>
<td>10</td>
<td>29</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>(3.7%)</td>
<td>(1.5%)</td>
<td>(2.2%)</td>
<td>(2.4%)</td>
</tr>
<tr>
<td>garbage</td>
<td>299</td>
<td>322</td>
<td>656</td>
<td>1277</td>
</tr>
<tr>
<td></td>
<td>(47.8%)</td>
<td>(49.2%)</td>
<td>(50.5%)</td>
<td>(49.5%)</td>
</tr>
<tr>
<td>doubtful</td>
<td>13</td>
<td>14</td>
<td>51</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>(2.1%)</td>
<td>(2.1%)</td>
<td>(3.9%)</td>
<td>(3.0%)</td>
</tr>
<tr>
<td>tot</td>
<td>626</td>
<td>654</td>
<td>1300</td>
<td>2580</td>
</tr>
</tbody>
</table>

Table 3: FIELD: statistics on the confirmation task of user B.

<table>
<thead>
<tr>
<th></th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>tot</th>
</tr>
</thead>
<tbody>
<tr>
<td>yu</td>
<td>101</td>
<td>139</td>
<td>941</td>
<td>1181</td>
</tr>
<tr>
<td></td>
<td>(44.3%)</td>
<td>(66.8%)</td>
<td>(70.4%)</td>
<td>(66.6%)</td>
</tr>
<tr>
<td>richy</td>
<td>72</td>
<td>17</td>
<td>109</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>(31.6%)</td>
<td>(8.2%)</td>
<td>(8.2%)</td>
<td>(11.2%)</td>
</tr>
<tr>
<td>garbage</td>
<td>38</td>
<td>39</td>
<td>176</td>
<td>253</td>
</tr>
<tr>
<td></td>
<td>(16.7%)</td>
<td>(18.7%)</td>
<td>(13.2%)</td>
<td>(14.3%)</td>
</tr>
<tr>
<td>doubtful</td>
<td>17</td>
<td>13</td>
<td>111</td>
<td>141</td>
</tr>
<tr>
<td></td>
<td>(7.5%)</td>
<td>(6.2%)</td>
<td>(8.3%)</td>
<td>(8.0%)</td>
</tr>
<tr>
<td>tot</td>
<td>228</td>
<td>208</td>
<td>1337</td>
<td>1773</td>
</tr>
</tbody>
</table>

**SPEECH RECOGNIZER**

The recognizer utilizes a set of phonetic units represented by continuous density Hidden Markov Models (HMMs). The acoustic features used are LPC cepstral coefficients and log-energy, with the corresponding first and second order time derivatives. RASTA filtering is applied for channel equalization. In order to model any word (for confirmation expressions) without losing accuracy on digit strings, we decided to duplicate the phones corresponding to digits. Table 4 shows the training sets of the various databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLEAN</td>
<td>5,215 sentences + ~30,000 digits</td>
</tr>
<tr>
<td>PHONE</td>
<td>1,473 sentences + 1,184 digit sequences + 423 confirmations</td>
</tr>
<tr>
<td>SIRVA</td>
<td>2,194 digit sequences</td>
</tr>
<tr>
<td>FIELD</td>
<td>507 digit sequences + 697 confirmations</td>
</tr>
</tbody>
</table>

Table 4: Total number of sentences used for training.

A previous work on the PHONE database, reported in [3], showed the importance of explicitly modeling some weak spontaneous speech phenomena. A large difference in Word Accuracy (WA) was observed when recognition was performed on the bigQ (97.21% WA) or on the medQ (85.33% WA) digit sequences, which contain those phenomena. To overcome this problem we introduced three new models: @eh, @br and @ns, representing hesitations, breaths and noises of various types, respectively. The HMMs of these new units were trained using the corresponding labeled occurrences in the PHONE database. Introducing them in the recognition network had no tangible effects on the bigQ test sentences, but performance on the medQ part raised to 95.81%. The global effect (on a mixed test set which contains both bigQ and medQ sentences) was a performance improvement from 94.75% WA to 96.95% WA.

The training procedure we adopt considers one word at a time, and performs Baum-Welch on the corresponding phone sequence. This means that a preliminary segmentation in words is needed, which was available both for CLEAN and PHONE databases. In order to obtain a reliable word segmentation on SIRVA and FIELD, we first performed a recognition (which also produces a segmentation in words) on each digit sequence. Then we retained only the digit
strings correctly recognized, following the reasonable assumption that, if the recognition is correct, the segmentation is sufficiently reliable\textsuperscript{1}. In this way a percentage of the training data is not used, but the segmentation is reliable without hand-checking.

Following [4], the HMM parameters are first initialized using CLEAN (hmm\textsubscript{1}), then PHONE is used to retrain models and to introduce the units @eh, @br and @ns (hmm\textsubscript{2}). Further retraining has been done using PHONE + SIRVA + FIELD (hmm\textsubscript{3}).

During recognition, a network which allows any combination of digits, background and the special units @eh, @br and @ns is used. For confirmation, a network representing both yn and all the expressions in richyn is used, in parallel with a rejection network composed by a subset of the phones. In additions, some other expressions explicitly represent garbage ("pronto" - "hallo", "chi parla" - "who is speaking", etc.). Different weights are applied to different paths of the network; their value either favours or penalizes garbages.

**RESULTS**

In table 5 recognition results are reported for two different test sets. The first one, PH+SI, consists of digit strings of PHONE and SIRVA (1465 sequences containing 5051 digits), the second one is composed by correct digit sequences (digits) of FIELD (488 sequences containing 4024 digits, with an average of about 9.5 digit per sequence). Performance on the FIELD test is also given in terms of Sentence Accuracy (SA). Furthermore, the various acoustic models, namely hmm\textsubscript{1}, hmm\textsubscript{2} and hmm\textsubscript{3}, are considered separately. As expected, performance improves by increasing the training material. The worst performance on FIELD is due, as previously observed, to the low quality of its signals. For this reason a spectral subtraction procedure [3] has been applied in order to increase the signal to noise ratio. However, as also observed in previous experiments on the PHONE database, no improvements have been obtained.

<table>
<thead>
<tr>
<th></th>
<th>PH+SI WA</th>
<th>FIELD WA</th>
<th>FIELD SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>hmm\textsubscript{1}</td>
<td>92.91%</td>
<td>93.51%</td>
<td>69.1% (337/488)</td>
</tr>
<tr>
<td>hmm\textsubscript{2}</td>
<td>95.51%</td>
<td>95.48%</td>
<td>76.2% (372/488)</td>
</tr>
<tr>
<td>hmm\textsubscript{3}</td>
<td>96.50%</td>
<td>95.96%</td>
<td>77.9% (380/488)</td>
</tr>
</tbody>
</table>

Table 5: Recognition results for digit sequences, using HMMs trained on increasing material.

As discussed before, for the confirmation task we did not consider the doubtful data, but only garbage, yn and richyn. Two test sets were defined, one concerning users A, the other concerning users B. Their sizes are reported in table 6. Evaluation was done by considering two parameters: the percentage of garbages detected (number of garbages correctly detected over total number of garbages) and the percentage of yes/no detected (number of yes/no correctly detected over total number of yes/no). The trade-off between yes/no detection and garbage detection can be controlled by means of weights in the network used by the recognizer. In this way a curve, shown in figure 2, can be drawn for users A and B.

<table>
<thead>
<tr>
<th></th>
<th>Users A</th>
<th>Users B</th>
</tr>
</thead>
<tbody>
<tr>
<td>yn + richyn</td>
<td>907</td>
<td>1223</td>
</tr>
<tr>
<td>garbage</td>
<td>955</td>
<td>214</td>
</tr>
<tr>
<td>total signals</td>
<td>1862</td>
<td>1437</td>
</tr>
</tbody>
</table>

Table 6: Test set for the confirmation task.

Users A present the worst performance; this is mainly due to fact that their signals are often affected by a high background noise, as many calls come from public boxes.

**ACKNOWLEDGMENTS**

We wish to thank G. Podda for having labeled part of the field data. We also thank CSELT for having provided the SIRVA database.

**References**


\textsuperscript{1}We retain also strings having only word insertions at the beginning or at the end, labeling these insertions as @ns.