Speaker Independent Recognition of Spontaneously Spoken Connected Digits


Abstract

- A system for automatic recognition of 10 digit merchant identification codes, and 15 digit customer credit card numbers, for the purpose of authorizing purchases charged to a credit card.
- 2000 connected digit strings from 1000 customers.
- The recognizer correctly recognized 97% of the digit strings.
- Recognition of the dollar amounts are presented with some preliminary results.

1. Introduction

- Evaluation of HMM under telephone network conditions, in a trial of a credit card authorization service (CAS).
- CAS task
  1) Merchant Identification (ID) number (a 10 digit number)
  2) Customer credit card number (a 15 digit number)
  3) Dollar amount of the sales transaction
- The system responds with an authorization code, if appropriate.

2. Speech Database

- The speech database was collected from the customers over 800-based
dialled up telephone lines.
- SONY EV-S800 VCR recorder digitized the incoming signal at 31.5 kHz and stored the 16 bit PCM data on a standard 8 mm video tape.
- If a spoken string had internal pauses of more than 200 msec, the individual pieces were edited and labeled (sub-string).
- A total of 2170 connected digit strings from 1000 talkers (about 2 strings/talker) were used in experiments.

3. Speech Recognition System
- 200–3500 Hz BPF, 8 kHz sampling rate.
- preemphasis $1 - 0.95z^{-1}$
- 45 msec Hamming window, 15 msec shift.
- 10-th order LPC
- 12-th order cepstrum and delta cepstrum
- 11 words + silence
- 2 models per word
- 8 states per model
- a maximum of 64 Gaussian mixture densities per state.

4. CAS Task Constraints on Digit Strings
- Finite state network (FSN) grammar. The average branching factor was reduced from 11 in the no-grammar case to 8. (ID and credit card numbers)
- The last digit of both the merchant ID number and the credit card number is a security digit called a checksum digit.
- Two ways of generating suboptimal strings
  1) multiple candidates
     - during the recognition process

2) alternate candidates
   - after the recognition process
   - A set of strings that simultaneously conforms to the grammar and satisfies the checksum criterion is generated from the top recognition candidate.

5. Multiple Candidates
   - For every frame of the input speech, and for every grammar node, the top $N$ accumulated likelihood scores and information about the grammar arcs traversed are stored.
   - The first string in this ordered list that passes the checksum rule is chosen as the recognized string. If none of the candidates pass the checksum criterion, the input is rejected.
   - problems
     1) Since the segmentation is derived from the best path, this scheme does not provide the best score possible for each candidate except for the top candidate.
     2) Since multiple segmentations for the same grammar arc are not saved during the recognition process, the second choices at a node may not even be the best second choice.

6. Alternate Candidates
   - Whenever the top recognition candidate is invalid, we can trivially generate a set of strings that are valid in the grammar and which do satisfy the checksum criterion.
   - Once these alternate candidates are generated by the above scheme, we can score each string independently according to its best path. No back-tracking is required.

7. Comparison of multiple and alternate candidates

- The Alternate Candidate scheme corrects single digit errors provided the correct string has a reasonable score. It is also able to correct some strings with two digit errors.

- The Multiple Candidate scheme does not even guarantee that the correct string will be in the list of candidates. One of the advantages of the Multiple Candidate approach is the ability to reject a string if a candidate with a valid checksum is not found in the list.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Wordspotting recognition string accuracy (in %) on independent data with extraneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No grammar</td>
</tr>
<tr>
<td>(Baseline) without checksum</td>
<td>82.0</td>
</tr>
<tr>
<td>Merchant ID</td>
<td>86.0</td>
</tr>
</tbody>
</table>

8. Natural Number Recognition

- Recognition of the dollar amount data is a much more difficult task than digit recognition because of its unconstrained nature and because its 37-word vocabulary of natural numbers includes many highly confusable words.
- HMM
  1) left to right HMM.
  2) single model per word + silence model.
  3) a maximum of 9 or 11 states per model.
  4) 16 Gaussian mixture densities per state.
- Data Base
  1) a training set of 6735 strings (4307 strings from 76 speakers, 2428 strings from 2428 speakers).
  2) a test set of 887 strings from 17 speakers.
  3) dialed up telephone lines.
- tied state HMM
  ; They generated context dependent subword models for the -ty, -teen extensions and concatenated these models to their root digit model.
- Higher order spectral and energy features

![Diagram of HMM](image)

Fig. 1. Tied state HMM.
Table 2  
Natural number string recognition accuracy (in %)

<table>
<thead>
<tr>
<th>No of states</th>
<th>HMM</th>
<th>Tied states HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>87.4</td>
<td>93.5</td>
</tr>
<tr>
<td>11</td>
<td>88.3</td>
<td>94.2</td>
</tr>
</tbody>
</table>

Table 3  
String recognition accuracy (in %) with improved acoustic resolution

<table>
<thead>
<tr>
<th>Features</th>
<th>DCEP (16 mix)</th>
<th>DDCEP +DDENG (16 mix)</th>
<th>DDCEP +DDENG (64 mix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of features</td>
<td>24</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>Recognition accuracy independent</td>
<td>83.9</td>
<td>88.6</td>
<td>90.2</td>
</tr>
</tbody>
</table>

9. Summary

- The raw error rates of the basic ASR algorithm with wordspotting capabilities were 1–2% per digit.
- By incorporating syntactic and semantic constraints, the overall system digit error rate was reduced to 0.2–0.3% per digit.
- The best system performance was 97.0% string accuracy for the 10 digit merchant numbers and 97.4% string accuracy for the 15 digit credit card numbers.