LANGUAGE MODELS FOR A SPELLED LETTER RECOGNIZER

Martin Betz and Hermann Hild

Interactive Systems Laboratories
University of Karlsruhe — 76128 Karlsruhe, Germany
Carnegie Mellon University — Pittsburgh, USA

ABSTRACT

In some speech recognition applications, it is rea-
e to constrain the search space of a speech rec-
rge but finite set of sentences. We demon-
a spelling task, where the recog-
ed last names is constrained
que names) of a tele-
ss this prob-
ge

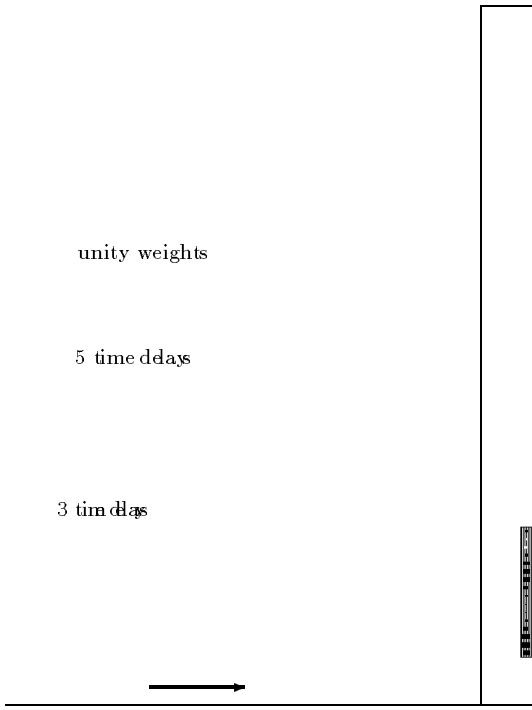

Figure 1: The MS-TDNN recognizing t
word 'B'. Only the activations for
are shown.

bel-
e plates"), "re-
for interactive recogniz-
names or addresses. In the latter
ries the search space can be constrained to
large dictionary of words or names. Constraints can
become effective *within the search process* as n-grams
or in a fully constrained search. They also can be used
to *postprocess* the recognized hypotheses by mapping
them onto legal strings, or by finding the highest rank-
ing legal hypothesis in an n-best list. In this paper, we
will demonstrate our letter recognizer and the effects
of various language models and search techniques on
the task of spelled name recognition. Related work on
isolated letters was reported by Cole et. al. [2].

classifier. Figure 1 s
of recognizing
16 mels

2. THE LETTER RECOGNIZER

The Multi-State Time Delay Neural Network (MS-TDNN) [3, 5] integrates the time-shift invariant architecture of a TDNN and a nonlinear time alignment procedure (DTW) into a high accuracy word-level

3. EXPERIMENTAL SETUP

e "Telephone Directory" used to constrain the search space contained 111,882 entries, with a total of 32,267 unique last names. After accounting for multiple pronunciation alternatives of some letters, the final list of names contained 43,181 strings, referred to as the string set $S = \{s_1, s_2, \ldots\}$. The recognizer was trained with 8,133 strings (55,449 letters) spelled by 70 speakers. The test set consists of 1,316 strings $\in S$ (8,661 letters) spelled by 23 additional speakers. S were sampled at 16 kHz with a Sennhei microphone. Except for th same test setup was u

4. I

As a baseline experiment, without any langua any other i

in the st match):

$$h^* = \begin{cases} h_1 \\ h_{i^*} \end{cases}$$

The recogniti the n-best list. For 85% Saturation occurs at a string accuracy, as shown in figure 2. (60.7%), the first-best choice matches an entry i list. About 5% of these first-best choices are incorrect. In 5.1% of all cases, none of the n hypotheses has a match in the dictionary. As expected, the percentage of misrecognitions increases as the first match occurs further down the n-best list. More detailed statistics are shown in table 3.

position	%	correct	incorrect
1	60.7	763	36
2	10.5	130	8
3	5.1	54	13
4	2.4	28	4
5	1.2	15	1
6	1.1	14	1
7	1.3	11	6
8	1.0	11	2
9	0.8	9	1
10	1.1	10	5
11 - 20	3.6	35	13
21 - 30	2.1	19	8
31 - 40	1.3	11	6
41 - 50	0.7	7	2
51 - 60	0.7	7	2
61 - 70	0.5	5	2
71 - 80	0.5	4	3
81 - 90	0.1	1	0
91 - 100	0.2	1	1
none	5.1	0	67

Table 3: The histogram shows with which frequencies
he best matching hypothesis was found at various po-
he n-best list.

LLY CONSTRAINED SEARCH

nstraints were applied *after*
ated; in this section
straints di-
inite

the letter E occurs in over 5,800 transitions. Since
the full left context of a string is considered during
the search, each transition may have a diffe
dividual accumulated search s
conventional or
sear

word in
in the minFSG graph,

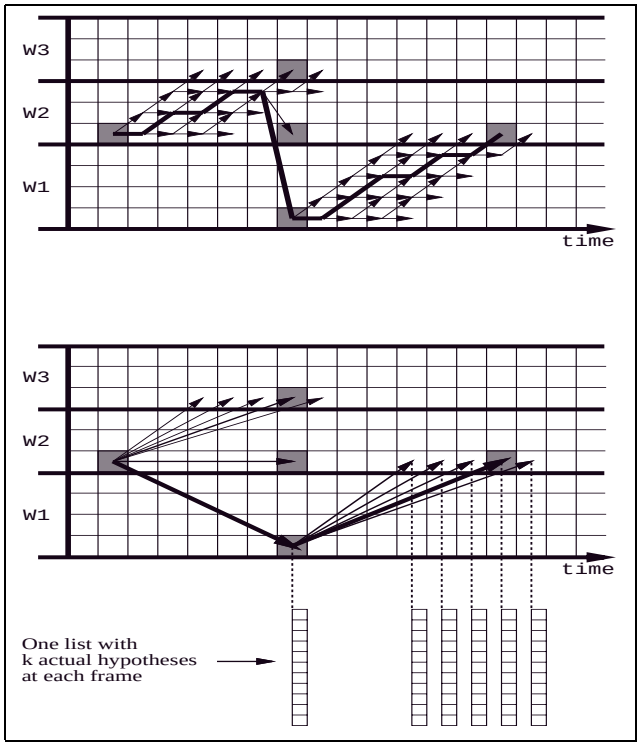

Figure 5: **Top:** Conventional DTW search technique. The matrix contains the prohibitive amount of 57,713 word models, one for each letter in the minFSG. **Bottom:** Two level search with only one word model for each letter in the alphabet, but an exp putation of partial scor the activ