European Conference on Speech, Communication and Technology (EUROSPEECH 93), Berlin, Germany,
Volume 2, pp. 1481-1484

SPEAKER-INDEPENDENT CONNECTED LETTER RECOGNITION WITH A MULTI-STATE TIME DELAY NEURAL NETWORK

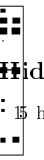
Hermann Hild and Alex Waibel

Universität Karlsruhe, Germany
Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

We present a Multi-State Time Delay Neural Network (MS-TDNN) for speaker-independent, connected letter recognition. Our MS-TDNN achieves 98.5/92.0% word accuracy on speaker dependent/independent English letter tasks[7, 8]. In this paper we will summarize several techniques to improve (a) continuous recognition performance, such as sentence level training, and (b) phonetic modeling, such as network architectures with "internal models", allowing for "tuning-in" to new speakers. We present results on our large and still growing letter data base, containing over 40.000 letters spelled by 55 speakers.

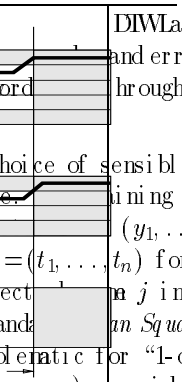
Letter Recognition, Speaker-



ION

of letters is essential
Input Layer
vocabularies, such as
small vocabulary, the
English or German let-
need further
similar sounds of (for
D and T. Throughout
er" and "word" in-
s to a string of

integrates the time-
[12] and a nonlin-
a word-level
process
by 16



DIWLayer. Instead of phonemes, the output are now
and error derivatives are backpropagated fromthe
ord through the alignment paths and the front-end

hoice of sensible objective functions is of great
e. ining on the phoneme level, there is
$(y_1, \ldots, y_n)$ and a corresponding
$= (t_1, \ldots, t_n)$ for each frame in time. T
ect me j in a "1-out-of-n" cod-
and *an Square Error* ($MSE =$
blematic for "1-out-of-n" codings
ur case); consider for example
nd
0), the output $(0.0, \ldots, 0.0)$
the more desirable output
pes
avoided by

$$- (y_i - t_i)^2)$$

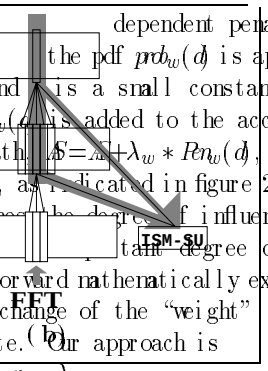
tliers" with an er-
proaching 1.0.

hieved best re-
he "Classifica-
o maximize
y_i and
ion)
solute
rts:

A B)
A A B)

A proper treat-
portant for a
at word
ce level
ord

naliz ing ofInser-
on Errors. In continuous recognition
ad of looking at word units the well-known
algorithm[11] is used to find an opti-
specified sequence of words. The
glishletters cause many word
such as "**T E**" vs. "**T**"
e proper duration model-
[6], minimumphoneme
plication". In ad-
and word (w)



dependent penalty $Pen_w(d) = log(k + prob_w(d))$, where the pdf $prob_w(d)$ is approximated from the training data nd k is a small constant to avoid zero probabilities. $_w(d)$ is added to the accumulated score AS of the th, $AS = AS + \lambda_w * Pen_w(d)$, whenever a word as indicated in figure 2(b). The ra- he degree of influence of the tant degree of free- orward mathematically exact change of the "weight" e. Our approach is nges. λ_w propor- he internal speaker re trying ture. Speaker letion

he Sentence Level.

on the en- trained to classify excerpted applied. ntinuously spoken sentences. utom- method to extend train- d shows the alignment hich a typical er- a forced alignment enforced), pos- path, while the tive training. training is ly differing

both

ker

DELS

letters

Models" (ISMs) is ak- e was measured work, each of which is spe- data, er, or a group of speakers, k- /female speakers. The vary from simple er-specific sub-

Data)	
NNs. When ces/words nor- our TDNN	
1	
6	
. 8	
86.9	
91.0	

(Res. Man. Spell-mode)	
000 train, 11/900 test speaker/words	
	our MS-TDNN
	gender specific
90.8	92.0

(in % on the test set) on speaker nected letter tasks.

e p e n d e n t Ge r m a n S p e l l i n g

4.2. GERMAN LETTERS

43/34823 train, 12/8206 test speaker/words

... Size | ... Word ... | ...
ayer) | train | test | train | test
its | 97.6 | 92.5 | 94.3 | 88.1
ts | 98.1 | 93.8 | 96.4 | 90.3

3: Word accuracy in % on German Spelling

We are in the process of continuously collecting a large data base of German spelled letters. At this time, more than 40.000 letters from 52 speakers (table 2) were collected and labeled. Volunteers were asked to spell a set of 50 to 150 sentences in a natural manner, without artificial pauses between letters. Each individual spells a different set, consisting of three categories: proper names, drawn randomly from a german list of 100.000 names, some random city names and pseudo-random letter sequences. The latter subset ...ed to increase the percentage of the less frequent ...h as **Q** or **X**, to make sure there is a reasonable ...ning data for all letters. For example, after ...udo' random sequences,' the ratio of **Q** to ...3 : 1000 to 75 : 1000 .

...gements

...ratefully acknowledge support by the NSF, ... Creating the large German data... ...le without the tireless work ...her, Henrike Triess, ...

...and phoneme boundaries, the data ...ng the JANUS LVQ recognizer ... After a initial training ...s used to relabel the ...he error rate on

...Systems

		all	
	Spr	letter	
	43	**34823**	
1865	**12**	**8206**	
3	11936	55	43029

...ell Data Base.

...addition to "Silence" and the ...lphabet, the German alpha...els ("Umlaute") **Ä**, **Ö**, ...ely, there are several ...the official version "es-...", "scharfes-S", or in ...also used. Since con-...en (**-**) was also ...erent possible ..."Gedanken-...elling