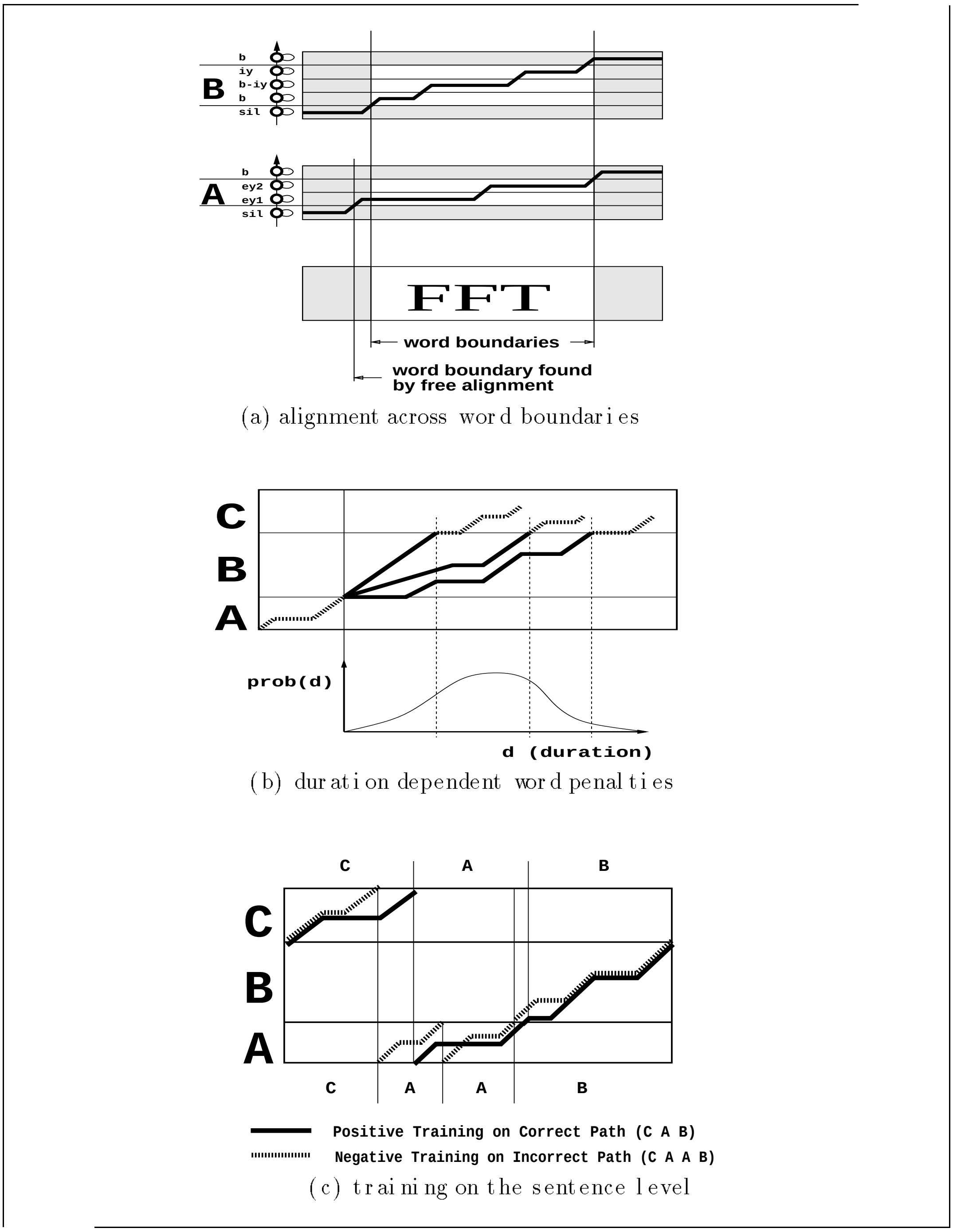

Figure 2: Various techniques to improve sentence level recognition performance

Speaker Dependent (CMU Alph Data)			
500/2500 train, 100/500 crossvalidation, 400/2000 test sentences/words			
speaker	SPHINX[HFW91]	MS-TDNN[HFW91]	our MS-TDNN
mjmt	96.0	97.5	**98.5**
mdbs	83.9	89.7	**91.1**
maem	–	–	**94.6**
fcaw	–	–	**98.8**
flgt	–	–	**86.9**
fee	–	–	**91.0**
Speaker Independent (Resource Management Spell-Mode)			
109 (ca. 11000) train, 11 (ca. 900) test speaker (words).			
SPHINX[HH92]		our MS-TDNN	
	+ Senone		gender specific
88.7	90.4	**90.8**	**92.0**

Table 1: Word accuracy (in % on the test sets) on speaker dependent and speaker independent connected letter tasks.

In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Toronto, Ontario, Canada, May 1991. IEEE.

[FC90] M. Fanty and R. Cole. Spoken letter recognition. In *Proceedings of the Neural Information Processing Systems Conference NIPS*, Denver, November 1990.

[Haf92] P. Haffner. Connectionist Word-Level Classification in Speech Recognition. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1992.

[HFW91] P. Haffner, M. Franzini, and A. Waibel. Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*. IEEE, 1991.

[HH92] M-Y. Hwang and X. Huang. Subphonetic Modeling with Markov States - Senone. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I33 – I37. IEEE, 1992.

[HW90] J. Hampshire and A. Waibel. A Novel Objective Function for Improved Phoneme Recognition Using Time Delay Neural Networks. *IEEE Transactions on Neural Networks*, June 1990.

P. Haffner and A. Waibel. Multi-state Time Delay Neural Networks for Con-
Speech Recognition. In *NIPS(4)*. Morgan Kaufman, 1992.

Waibel. Multi-Speaker/Speaker-Independent Architectures for
me Delay Neural Network. In *Proc. IEEE International*
s, Speech, and Signal Processing. IEEE, 1993.

and R. A. Cole. Speaker-independent Phonetic
lish Letters. In *Proceedings of the IJCNN*

mic Programming Algorithm for Con-
ons on Acoustics, Speech, and
984.

K. Lang. Phoneme
sactions on

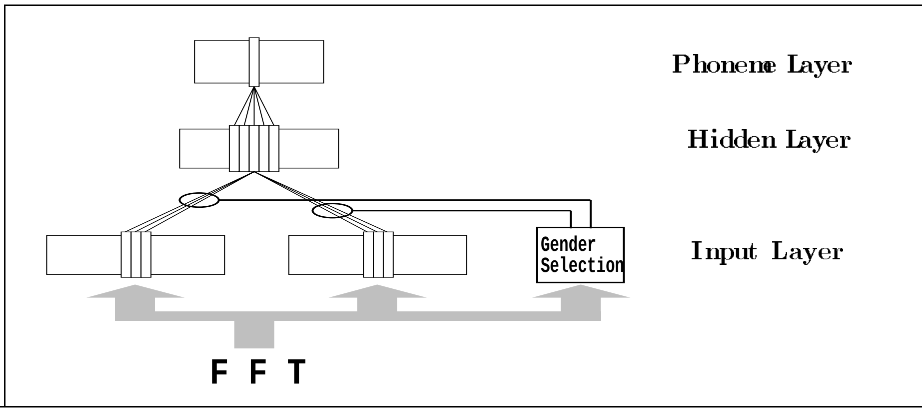

Figure 4: A network architecture with gender-specific and shared connections. Only the front-end TDNN is shown.

ing, cross-validation and test set, respectively. The DARPA Resource Management Spell-Mode Data were used for **speaker independent** testing. This data base contains about 1700 sentences, spelled by 85 male and 35 female speakers. The speech of 7 male and 4 female speakers was set aside for the test set, one sentence
from all 109 and all sentences from 6 training speakers were used for crossvali
Table 1 summarizes our results. With the help of the training techni
above we were able to outperform previously reported [HFW91] s
results as well as the HMM based SPHINX System.

5 SUMMARY AND FUTURE WORK

We have presented a connectionist speech recognition s
connected letter recognition. New training techni
level recognition enabled our MS-TDNN to out
kind as well as a state-of-the art HMM bas
specific subnets, we are experimenting
"internal speaker models" for a more
the future we will also experi

Acknowledgements

The authors gratefully ackn
and DARPA. We wish to t
McNair for keeping o
the ideas presen

Refe re

[CFGJ91]
In

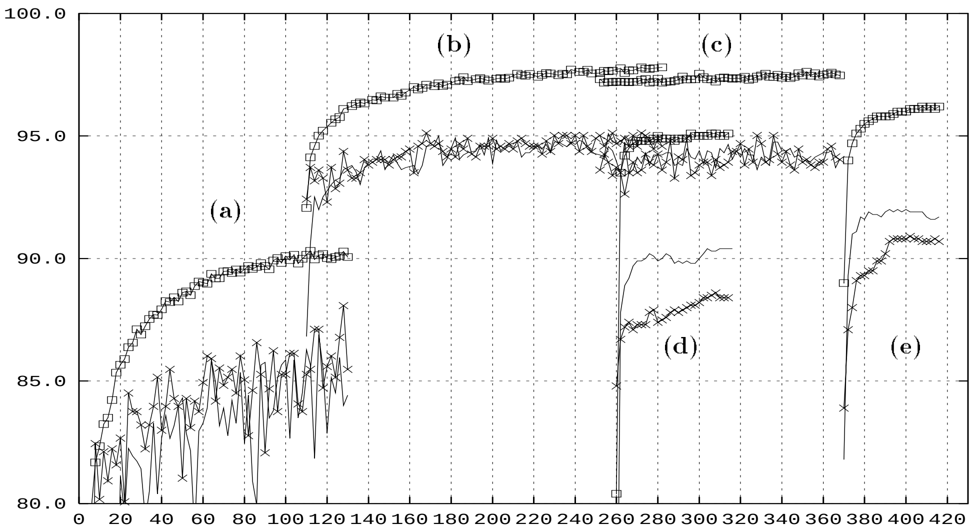

Figure 3: Learning curves (a = bootstrapping, b, c = word level (excerpted words), d, e = sentence level training (continuous speech)) on the training (□), crossvalidation (-) and test set (x) for the speaker-independent RM Spell-Mode data.

3 GENDER SPECIFIC SUBNETS

A straightforward approach to building a more specialized system
two entirely individual networks for male and female speake
the gender of a speaker is known, during testing it
"gender identification network", which is si
put units representing male and female sp
network classifies the speaker's gen
ularized network improved the
see table 1) to 91.3% How
connections at the
worked even
cati

in the same way as the phoneme boundaries within a word. Figure 2(a) shows an
example in which the word to recognize is surrounded by a silence and a 'B', thus
the left and right context (for all words to be recognized) is the phoneme 'sil' and
'b', respectively. The gray shaded area indicates the extension necessary to the
DTW alignment. The diagrams shows how a new boundary for the beginning of
the word 'A' is found. As indicated in figure 3, this techniques improves conti
recognition significantly, but it doesn't help for excerpted words.

2.2 WORD DURATION DEPENDENT PENALIZING OF INSERTION AND DELETION ERRORS

In "continuous testing mode", instead of looking at word units the well-known "One
Stage DTW algorithm[Ney84] is used to find an optimal path through an unspec-
ified sequence of words. The short and confusable English letters cause many word
insertion and deletion errors, such as "T E" vs. "T" or "O" vs. "O O", ther
proper duration modeling is essential.

As suggested in [HW92], minimum phoneme duration can be enforce
duplication". In addition, we are modeling a duration and word de
$Pen_w(d) = log(k + prob_w(d))$, where the pdf $prob_w(d)$ is app
training data and k is a small constant to avoid zero p
added to the accumulated score AS of the search p
whenever it crosses the boundary of a word w in
rithm, as indicated in figure 2(b). The ratio λ
influence of the duration penalty, is anot
is no straightforward mathematical
of the "weight" λ_w to the inser
gradient descent, which c
i.e. we are trying to n

2.3 ERROR I

Usually the MS-TDNN is trained to
tinuously spoken sentences.
training on the sentence
"C A B", in whi
alignme

copied from the Phoneme Layer into the word models of the DTW Layer, where an optimal alignment path is found for each word. The activations along these paths are then collected in the word output units. All units in the DTW and Word Layer are linear and have no biases. 15 (25 to 100) hidden units per frame were used for speaker-dependent (-independent) experiments, the entire 26 letter network has approximately 5200 (8600 to 34500) parameters.

Training starts with "bootstrapping", during which only the front-end TDNN
used with fixed phoneme boundaries as targets. In a second phase, trainin
formed with word level targets. Phoneme boundaries are freely align
word boundaries in the DTW layer. The error derivatives are ba
the word units through the alignment path and the front-en
The choice of sensible objective functions is of gre
$(y_1, \ldots, y_n)$ the output and $T = (t_1, \ldots, t_n)$ th
the phoneme level (bootstrapping), ther
time, representing the correct pho
see why the standard *Mean*
for "1-out-of-n" codin
for a target (1.
th

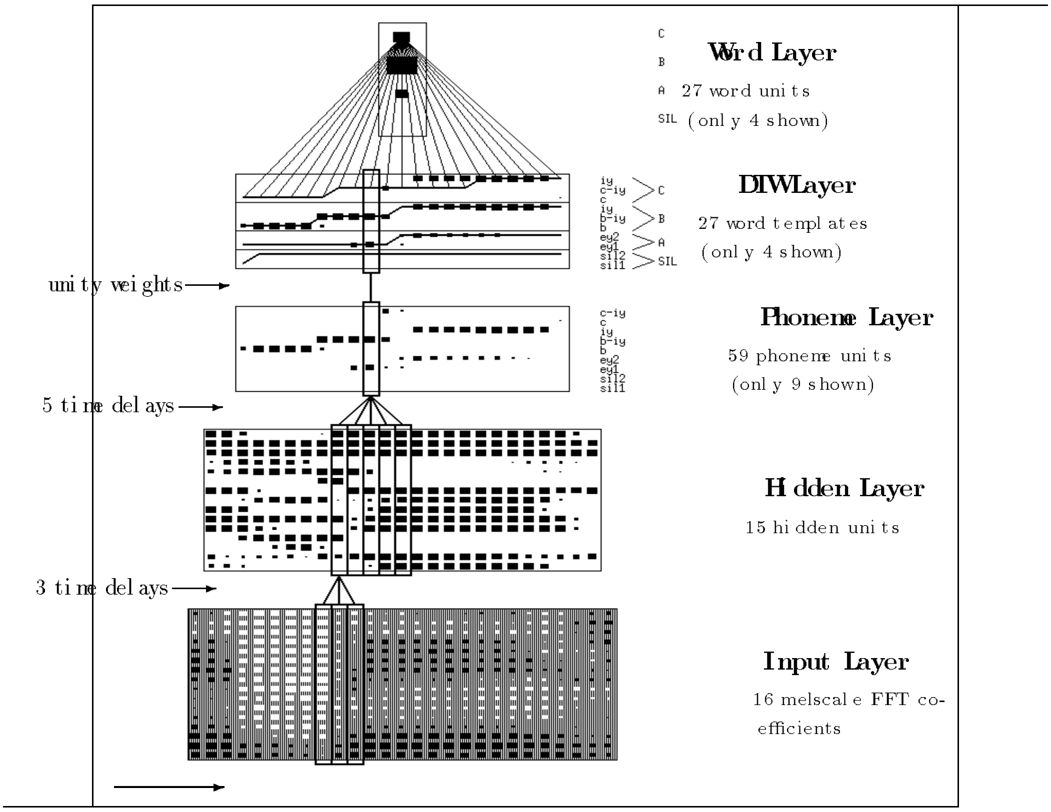

Figure 1: The MS-TDNN recognizing the excerpted word 'B'. Only the activations for the words 'SIL', 'A', 'B', and 'C' are shown.

classified by another network. In this paper, we present the MS-TDNN as a connectionist speech recognition system for connected letter recognition. After describing the baseline architecture, training techniques aimed at improving sentence level performance and architectures with gender-specific subnets are introd

Baseline Architecture. Time Delay Neural Networks (TDNNs) can comb robustness and discriminative power of Neural Nets with a time-shift chitecture to form high accuracy phoneme classifiers [WHH+89] TDNN (MS-TDNN) [HFW91, Haf92, HW92], an extension of the TI of classifying words (represented as sequences of phonemes) by ear time alignment procedure (DTW) into the TDNN archi an MS-TDNN in the process of recognizing the excerpt 16 melscale FFT coefficients at a 10-msec frame ra tute a standard TDNN, which uses sliding w to compute a score for each phoneme (st tions in the "Phoneme Layer". In is modeled by a sequence of pho

*Advances in Neural Information Processing Systems 5 (NIPS*5).*
San Mateo, CA: Morgan Kaufmann Publishers, 1993

Connected Letter Recognition with a Multi-State Time Delay Neural Network

Hermann Hild and **Alex Waibel**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3891, USA

Abstract

The Multi-State Time Delay Neural Network (MS-TDNN) integrates a nonlinear time alignment procedure (DTW) and the high-accuracy phoneme spotting capabilities of a TDNN into a connectionist speech recognition system with word-level classificatio
error backpropagation. We present an MS-TDNN for recog
continuously spelled letters, a task characterized b
highly confusable vocabulary. Our MS-TDNN achie
word accuracy on speaker dependent/independe
forming previously reported results on the
pose training techniques aimed at impr
mance, including free alignment
ration modeling and error ba
than the word level. Arc
ized on a subset

1 INTROD

The recogni