*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 speaker models", allowing for "tuning-in" to new speakers. We also present results on our large and still growing German Letter database, containing over 40.000 letters continuously spelled by 55 speakers.

elled Letter Recognition, Speaker-
ce, MS-TDNN

## 1. INTRODUCTION

ition of spelled strings of letters is essential for all application involving special vocabularies, such as names or addresses. Despite its small vocabulary, the task is quite difficult because the English or German letters are easily confused. Even humans often need further inquiry to distinguish between the similar sounds of (for example) the letters **M** and **N**, or **D** and **T**. Throughout this text, we will use the terms "letter" and "word" interchangeable. The term "sentence" refers to a str letters.

**The Baseline MS-TDNN** [5, 8] inte shift invariant architecture of ear time alignment proc classifier. Figur of recogniz nel

DTW Layer. Instead of phonemes, the output are now words, and error derivatives are backpropagated from the word units through the alignment paths and the front-end TDNN.

The choice of sensible objective functions is of great importance. For training on the phoneme level, there is an output vector $Y = (y_1, \ldots, y_n)$ and a corresponding target vector $T = (t_1, \ldots, t_n)$ for each frame in time. $T$ represents the correct phoneme $j$ in a "1-out-of-$n$" coding, i.e. $t_i = \delta_{ij}$. Standard *Mean Square Error* ($MSE = \sum_{i=1}^{n}(y_i - t_i)^2$) is problematic for "1-out-of-$n$" c for large $n$ ($n > 50$ in our case); consider that for a target $(1,0,\ldots,0)$, the has only half the error tha $(1.0, 0.2, \ldots, 0.2)$. Thi

*McClell an*

which (like cros ror a

dependent penalty $Pen_w(d) = log(k + prob_w(d))$, where
the pdf $prob_w(d)$ is approximated from the training data
is a small constant to avoid zero probabilities.
led to the accumulated score $AS$ of the
$* Pen_w(d)$, whenever a word
re 2(b). The ra-
f the

## 4.2. GERMAN LETTERS

We are in the process of creating a large data base of German spelled letters. At this time, more than 40.000 letters from 55 speakers (table 2) were collected and labeled.
nteers are asked to spell a set of 50 to 150 sentences
ner, without artificial pauses between let-
a different set, consisting of
randomly from a