SEE ME, HEAR ME: INTEGRATING AUTOMATIC SPEECH RECOGNITION AND LIP-READING

Paul Duchnowski[1] *Uwe Meier*[1] *Alex Waibel*[1,2]

[1] University of Karlsruhe, Karlsruhe, Germany
[2] Carnegie Mellon University, Pittsburgh PA, USA

ABSTRACT

t recent work on integration of visual information acoustic speech for bet-

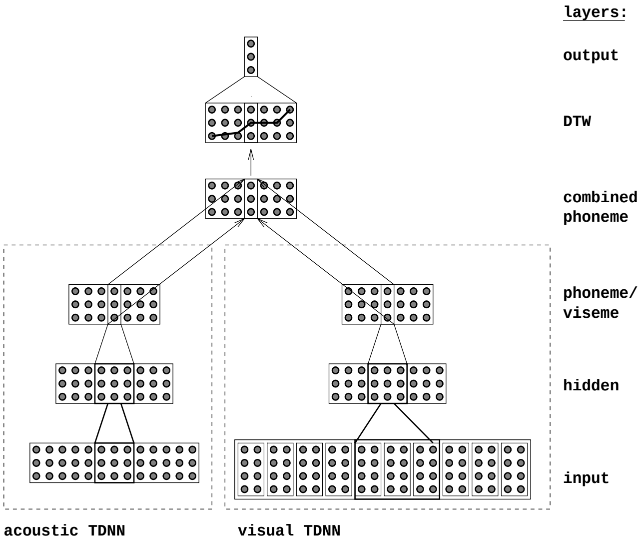


Figure 1. Original recognition network architecture (Net-P).

data, respectively. Weighted sums of the phone and corresponding viseme activations are entered in the combined layer and a one stage DTW algorithm finds the optimal path through the phone states that decodes the recognized letter sequence. The weights in the parallel networks are trained by backpropagation. There are 15 hidden units in both subnets. The combination weights are computed dynamically during recognition to reflect the estimated reliability of each modality. These "entropy weights" [2], λ_A for the acoustic side and λ_V for the visual are given by:

$$\begin{aligned} \lambda_A &= b + \frac{S_V - S_A}{\Delta S_{max-over-data}} \\ \lambda_V &= 1 - \lambda_A \end{aligned} \qquad (1)$$

The entropy quantities S_A and S_V are computed for the acoustic and visual phone/viseme activations by normalizing these to sum to one and treating them as probability mass functions. High entropy is found when activations are evenly spread over the units which indicates high ambiguity of the decision from that particular modality. The bias b pre-skews the weights to favor one of the modalities.

2.2. Visual Data Representation

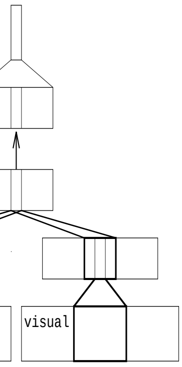
Unlike for acoustic speech data, there are no generally agreed-upon parameterization strategies for the visual lip images. Since we are using a connectionist algorithm for recognition we have followed the philosophy of avoiding ex-
feature extraction and segmentation of the image. In-
on the network to develop appropriate inter-
higher level features. We have been
e visual data representations

ixel vector is quite
input vec-
o

Visual Input	Parameter Count	Word Accuracy (%) data set mm1-2	mm9-10
Gray Levels	384	55	44
	32	52	45
	32	53	52
	29	50	38

y recognition rates for different data repre-

ents the recognizer from taking advantage of
ons between acoustic and visual events
ationships. There is evidence
nputs to take advan-



Net H

ic and visual combi-

srated in Figure 2.
be con-

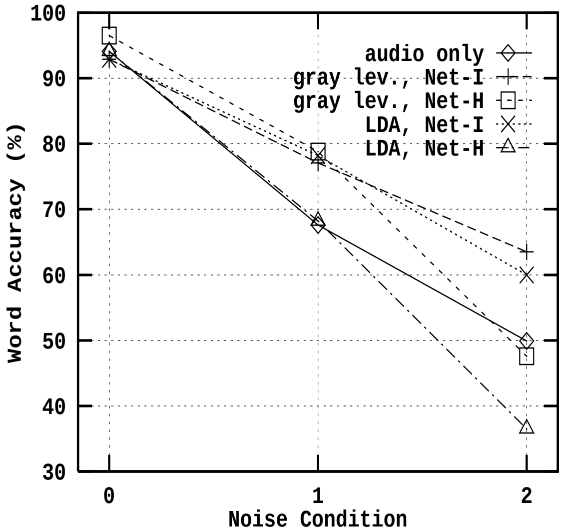


Figure 4. Combination results for Net-I and Net-H.

Comparison of different net structures yields more equiv-
ocal conclusions. All three are clearly capable of improving
recognition with the addition of visual information. How-
Net-P combination of the modalities *always* yields a
than either modality alone which is not true of
res. On the other hand, neither Net-I
at this time (for instance,
rited from Net-P).
lently for