A Novel Objective Function for Improved Phoneme Recognition Using Time Delay Neural Networks

J. B. Hampshire II A. H. Waibel

Department of Electrical & Computer Engineering and School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213-3890

Abstract

We present single and multi-speaker recognition results for the voiced-stop consonants /b, d, g/ using Time Delay Neural Networks (TDNN) [1], a new objective function for training these networks, and a simple arbitration scheme for improved classification accuracy. With these enhancements we achieve a median 24% reduction in the number of misclassifications made by TDNNs trained with the traditional backpropagation objective function. This reduction results in /b, d, g/ recognition rates that consistently exceed 98% for TDNNs trained with individual speakers; it yields a 98.1% recognition rate for a TDNN trained with three male speakers.

1 Introduction

TDNN architectures have been applied to the task of voiced-stop consonant phoneme recognition with excellent results [1,2]. In moving from speaker-dependent phoneme recognition to speaker-independent recognition, we consider a collection of enhancements to the TDNN that yields improved single and multi-speaker recognition results for the /b, d, g/ phoneme recognition task. These enhancements entail the development of an alternative objective function for the n-dimensional gradient search of backpropagation learning. We term this new objective function the "classification figure-of-merit" (CFM) in reference to the emphasis it places on the classification result obtained from the network. This emphasis differs markedly from that of the traditional mean-squared-error (MSE) function, which explicitly seeks to match the network's output with some ideal target (the notion of "correct classification" is only implicit in this objective). Our preliminary results show equivalent quantitative performance for MSE and CFM classifiers, but markedly different qualitative performance. Specifically, the different objective functions produce equivalent recognition performance, yet they engender substantially different feature abstractions, resulting in largely disjoint misclassified tokens. Using a simple arbitration mechanism, one can exploit this qualitative difference to reduce by 24% the number of misclassifications made by the MSE classifier alone[1]. We call this mechanism "conflict arbitration". The arbitration process identifies or "flags" 70% of the post-arbitration misses, at the cost of flagging 8% of the post-arbitration hits (9% of the entire token set) as possible misses. These techniques result in single-speaker /b, d, g/ recognition rates that consistently exceed 98%. Additionally, they achieve a 98.1% recognition rate for a TDNN trained with three male speakers.

The experimental conditions under which these findings were made are detailed in [1] and [2]. Japanese speech data was obtained from six professional announcers (4 male, 2 female), sampled at 12 kHz, parsed for the /b, d, g/ phonemes, and hamming windowed; from this windowed data 256-point DTFTs were computed at 5 msec intervals. The DTFTs were used to generate 16 Melscale coefficient spectra [1] at 10 msec intervals. These spectra were normalized to produce suitable input levels for the TDNNs. Training tokens for individual speakers were shuffled randomly and interleaved to produce successive /b, d, g/ tokens. Training tokens for TDNNs trained with multiple speakers were prepared similarly with the additional step of interleaving the tokens for a given phoneme across all speakers. Figure 1 illustrates the TDNN architecture trained with this data. The input layer comprises 15 16-coefficient Melscale spectra. TDNN connections between lower and higher layers of the network are linked in the time domain to engender shift-invariant pattern recognition. Details of this shift-invariant connectionist architecture can be found in [1]. In Figure 1 black indicates positive node activation, gray indicates no activation, and white (input layer only) indicates negative activation. A node's activation level is proportional to the size of the black or white rectangle depicting the node.

2 A review of the MSE objective function

In presenting the CFM objective function, we first review the traditional MSE objective function used in backpropagation [3,4]. This function seeks to minimize the mean-squared-error between the network's output nodes $\mathcal{O}_1, \ldots, \mathcal{O}_n$ and an ideal

[1] all statistics quoted in this paper are median values, owing to the small sample size (n=8).

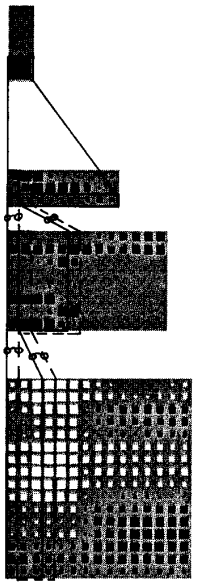

Figure 1: A Time-Delay Neural Network

or desired set of outputs $\mathcal{D}_1, \ldots, \mathcal{D}_n$

$$MSE = 1/n \sum_{i=1}^{n} (\mathcal{O}_i - \mathcal{D}_i)^2 \quad (1)$$

$$\mathcal{E} = 1/2 \sum_{i=1}^{n} (\mathcal{O}_i - \mathcal{D}_i)^2 \quad (2)$$

where (2) is the form used in [3,4] and $\mathcal{E} = MSE \cdot n/2$.

Clearly $\nabla_{MSE}(\mathcal{W})$ depends upon the conditional nature of the training token set. Indeed, one finds that the MSE objective function produces representations within the network that result from minimization of the *global* MSE (i.e., the MSE produced across all training tokens). We raise the question of whether the MSE objective function is optimal for training networks employed as classifiers.

Generalization is a term with broad implications in connectionist learning. For the purpose of our presentation, we address one aspect of its meaning for networks employed as classifiers. In this restricted context, generalization is a description of a network's ability to form abstract representations of a training set's salient features in order to maximize the number of correct classifications made on a disjoint test set. All the variables of a connectionist structure — the network architecture, its final connection strengths, the learning algorithm used to develop those connection strengths, and the statistical nature of the training set — play a role in determining the degree to which a network forms general representations. These variables also determine the specific *nature* of the resulting general representations. Reference [5] illustrates the importance of training set selection in the development of generalized representations, focusing on networks that deal with training patterns drawn from a finite, deterministic ensemble. We suggest that the backpropagation objective function plays an equally important role in forming general representations — particularly in networks that analyze training sets drawn from an infinitely large, stochastic ensemble characterized by a high degree of variability.

For the case of a classifier network with n outputs representing n possible classes, one can show for a single input pattern that one is assured a hit (correct classification) iff $MSE < 1/2n$. Likewise, one is assured a miss (incorrect classification) iff $MSE > (n-1)/n$. These facts form the basis of the phenomenon commonly know as "over-learning" or "rote memorization", often displayed by backpropagation networks trained with the MSE objective function. In short, because the MSE objective function is minimized *globally*, one finds that when training tokens are selected from an infinitely large high-variance ensemble, MSE is minimized for the vast majority of the tokens, while it is made quite large for a small minority of tokens. Given the hit/miss boundaries above and the fact that these minority tokens are legitimate samples from the ensemble, one often finds that recognition performance on test data actually degrades as learning of the training set continues beyond some optimal stage. One typically attempts to prevent this sort of pathological behavior in one of two ways: 1) by expanding the training set, using its statistical variance to obscure features that are not representative of the ensemble; 2) by explicitly selecting the training set, choosing tokens that are most representative of the ensemble and most effective in developing optimal classification boundaries [5]. The first solution is often impossible because one does not have access to a sufficiently large sample set. The second solution requires a-priori knowledge of those features that are representative of the ensemble; this presents a paradox, since one is attempting to train the network to *find* these features. Additionally, the selection task becomes extremely complex for training sets drawn from a very large high-variance ensemble. We suggest the alternative solution of modifying the means by which the network measures its "knowledge" of a training set; given a fixed backpropagation network architecture, this means is expressed in the objective function.

3 The CFM objective function

The CFM objective function has three essential features that distinguish it from the traditional MSE objective function:

- It has no notion of an *ideal target* classification output *pattern* to which it should match its output. Instead, it is only concerned that the output node representing the correct classification outcome have a higher activation state than any other output node. Its continuous mathematical form assesses a measure of the degree to which the correct classification has or has not been made — a classification

figure-of-merit.

- In order to discourage the network from attempting to produce ideal output patterns (thereby tending toward specific rather than general representations of the training set), the objective function yields decreasing marginal "rewards" for increasingly ideal output patterns.
- In order to discourage the network from attempting to learn tokens that are extreme statistical outliers for their given class, the objective function yields decreasing marginal "penalties" for increasingly bad misclassifications.

The resulting CFM objective function first compares the activation level of the output node that should be at high state with the activations of all other nodes which, in a classifier, should be at low state. It then applies a sigmoidal function to each of these differences. In this way, learning focuses most heavily on the reduction of misclassifications, rather than on attempts to mimic a target output exactly:

$$CFM = \sum_{\substack{n=1 \\ n \neq T}}^{N} \frac{\alpha}{1 + e^{(-\beta \Delta_n + \zeta)}}, \tag{3}$$

where

$\Delta_n = \mathcal{O}_T - \mathcal{O}_n$
$\mathcal{O}_T \equiv$ the "true" (i.e., correct classification) node
$\mathcal{O}_n \equiv$ the non-true node n
$N =$ total number of classes
$\alpha \equiv$ sigmoid scaling parameter
$\beta \equiv$ sigmoid discontinuity parameter
$\zeta \equiv$ sigmoid lateral shift parameter

Thus,

$$\frac{\partial CFM}{\partial \mathcal{O}_n} = -\alpha\beta \cdot \mathcal{Y}_n(1 - \mathcal{Y}_N) \tag{4}$$

$$\frac{\partial CFM}{\partial \mathcal{O}_T} = \alpha\beta \cdot \sum_{\substack{i=1 \\ i \neq T}}^{n} \mathcal{Y}_n(1 - \mathcal{Y}_N), \tag{5}$$

where

$$\mathcal{Y}_n \equiv \frac{1}{1 + e^{(-\beta \Delta_n + \zeta)}}$$

Equations (3) through (5) are variants of the well-known sigmoid function and its derivative [3,4]. Clearly there are many other functions which meet the CFM specifications itemized above; we present this particular form as an archetype from which further developments might be made. As mentioned earlier, equations (3) through (5) form a mathematically continuous expression of the degree to which a classifier produces the desired output classification. Note again that this measure of degree — this figure-of-merit — emphasizes the *relative* activations of all output nodes rather than their correspondence

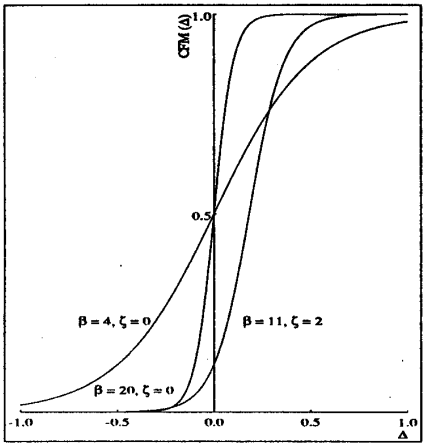

Figure 2: CFM plotted for representative parameter values

with some "ideal" output state. Because one seeks to maximize the CFM objective function, the weight-deflection equation of [3,4] must be changed to perform gradient ascent (as opposed to gradient descent):

$$\Delta W(t) = +\varepsilon \frac{\partial CFM}{\partial W(t)} + \alpha \Delta W(t-1) \tag{6}$$

Figure 2 illustrates the CFM function over the [-1,1] domain of Δ_n for some representative parameter values. The parameter β of (3) determines how discontinuous the sigmoid function is. As β becomes large, the CFM function approximates the Heaviside step function, and its derivative approximates the Dirac delta function. The β parameter allows one to modify the CFM function in terms of the degree of increasing marginal credit it assigns to an increasingly strong hit as well as the amount of increasing penalty it assigns to an increasingly strong miss. The ζ parameter sets the relative credit assigned to a classification that is on the borderline between a hit and miss (i.e., $\Delta_n \approx 0$ for some n) [2]. Our initial results show that the CFM classifier is quite responsive to changes in the β parameter. In particular, large values of β ($\approx$ 20) engender connection strengths that yield remarkably weak marginal hits with high MSE in both training and testing data, while smaller values of β ($\approx$ 4) yield strong hits that exhibit MSE comparable with that produced by the MSE classifier. Additionally, it appears that large values of β engender a diminished ability to discern subtle features necessary for high accuracy classification (manifest in reduced phoneme recognition rates). This is because the objective function is essentially flat for large values of Δ_n; as a result, it does not alter classification boundaries in response to more subtle features of the training set. Although the detailed effects of

[2] the α parameter is a simple scaling factor, typically equal to unity.

Network	Speaker	MSE	CFM	MSE/CFM
TDNN	MAU	98.3	98.9	98.8
	MHT	99.7	99.5	99.7
	MNM	97.4	97.2	98.3
	FKN	97.6	97.8	98.1
	FSU	98.2	98.5	98.4
	MMS	97.7	98.5	98.5
TDNN	1st 3	97.3	97.5	98.1
	all 6	95.9	95.9	96.5

Table 1: Comparison of /b, d, g/ recognition rates for TDNN trained with MSE, CFM and arbitrated MSE/CFM objective functions (CFM paramters: $\alpha = 1.0$, $\beta = 4.0$, $\zeta = 0.0$).

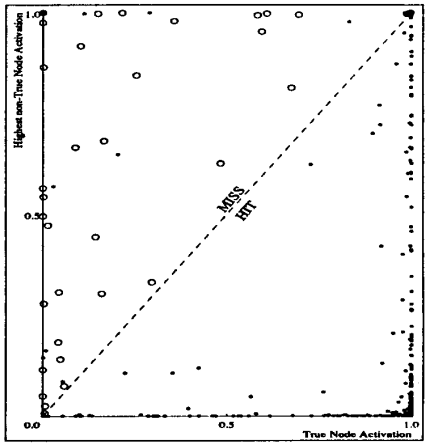

Figure 3: Scatter plot of MSE classifier outcomes. ○ indicates MSE miss correctly classified by CFM.

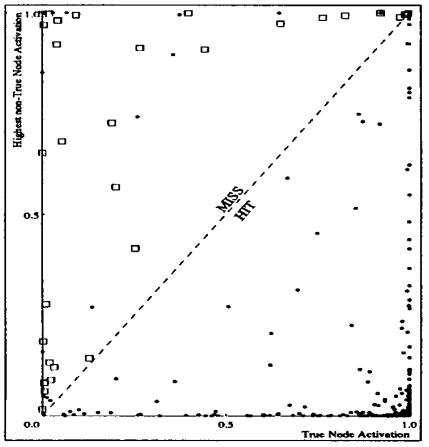

Figure 4: Scatter plot of CFM classifier outcomes. □ indicates CFM miss correctly classified by MSE.

altering ζ are not well-known we include it in (3) - (5) in order to provide a mechanism for specifying the relative magnitude of the CFM function for borderline tokens. In our preliminary studies, we have found that the parameter choices $\beta = 4.0$, $\zeta = 0.0$, and $\alpha = 1.0$[3] yield recognition rates for the /b, d, g/ recognition task that are equivalent to those for the MSE classifier. For $4 \leq \beta \leq 20$ we have found no evidence to suggest that the CFM function exibits the over-learning tendency of the MSE function — a definite advantage for the CFM function.

Table 1 shows the results of training a TDNN with tokens from 6 individual speakers, as well as two combinations of speakers using the MSE and CFM objective functions (see columns "MSE" and "CFM"). Recognition rates for MSE-trained TDNNs are based on training sessions monitored for the inception of over-learning. CFM recognition rates are based on unmonitored training sessions. Under these conditions, we find the recognition rates for the two classifiers equivalent.

[3]these parameter choices effectively reduce the CFM to a one-parameter function.

4 Conflict arbitration

The results of the previous section indicate the quantitatively equivalent recognition rates of the MSE classifier and the CFM classifier ($\beta = 4.0$, $\zeta = 0.0$, $\alpha = 1.0$) given a TDNN architecture. In this section we describe in general terms the qualitative differences between the two classifiers and the means by which these differences can be exploited for improved recognition rates.

In developing the CFM classifier, our principle goal was to produce a more appropriate objective function for connectionist classifiers; a by-product of this goal has been the development of an objective function that forms internal abstract representations of training tokens markedly different from those of the MSE classifier. A number of peripheral observations substantiate this assertion. First, we take a number of weight vectors for fully trained MSE TDNNs and use these as input weight vectors for CFM training sessions. The CFM classifier consistently evaluates these initial weight vectors as sub-optimal, yielding

final CFM weight vectors that are substantially different [4] from their MSE starting-points. Additionally, we consistently find that the set of MSE misses and the set of CFM misses are largely disjoint. The scatter plots of Figures 3 and 4 illustrate this phenomenon for all phonemes (b, d, and g) of the TDNN trained with 3 speakers, listed in Table 1. The results for this network are representative of those for the other networks in Table 1. Each plot shows the level of activation for the most active non-true output node (i.e. the most active node that does *not* represent the correct classification) versus the level of activation for the true output node (representing the correct classification). Thus, hits fall below and to the right of the dashed line and misses fall above and to its left. In both plots hits and misses common to both classifiers are shown as •. Figure 3 shows results for the MSE classifier and identifies those MSE misses that are CFM hits (○). Likewise, Figure 4 identifies CFM misses that are MSE hits (□). It is clear from both figures that the two classifiers have few common misses. In fact, if one considers the union of all missed tokens for the two classifiers, one typically finds that only 30% of these are common to both classifiers, while the remaining 70% are disjoint.

Using the results of Figures 3 and 4, and those of the other experiments listed in Table 1 one can develop a simple set of criteria to arbitrate a classification decision when the CFM and MSE classifiers disagree. In this way, on can reduce the number of misclassifications. The arbitration criteria used to choose between conflicting outcomes of the two classifiers can be stated qualitatively quite simply:

- if the CFM classifier is confident in its classification, choose the CFM outcome; do not flag the token as a possible miss unless the MSE classifier is also confident.
- otherwise, if the total output activation of both classifiers is weak, choose the MSE outcome and flag the token as a possible miss.
- otherwise, choose the more confident of the two outcomes; do not flag the token as a possible miss if the chosen classifier is far more confident than its competitor.

Note that the arbitration scheme flags tokens for which arbitration gives no clear choice. Beyond the arbitration process, flagging is also performed on tokens for which both classifiers agree but have substantially different activation levels. This provides a mechanism by which post-arbitration misses are often identified as suspect classifications. The confidence measure cited in the criteria above is computed using the CFM function applied to the outcome of each classifier under the assumption that the outcome is correct. Figure 5 illustrates the median 23.7% reduction in MSE misclassifications achieved through conflict arbitration. Comparing Figure 5 with Figures 3 and 4, one can see that the arbitration scheme is particularly effective in eliminating those misses that fall along the hit-miss

[4] it is not unusual to find the MSE and CFM weight vectors computed in this manner nearly orthogonal.

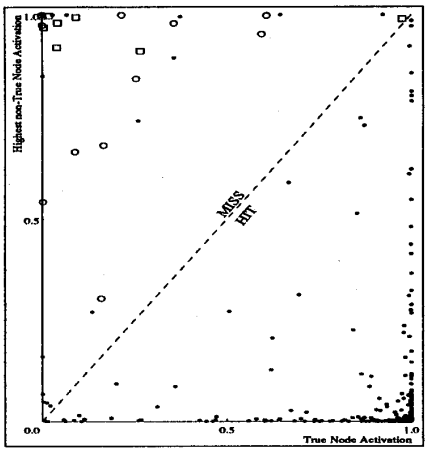

Figure 5: Scatter plot of arbitrated MSE/CFM classifier outcomes. □ indicates post arbitration miss correctly classified by MSE. ○ indicates post arbitration miss correctly classified by CFM.

borderline. Table 1 lists the improved recognition results for conflict-arbitrated MSE/CFM classification ("MSE/CFM" column). TDNNs trained with single-speaker data consistently yield recognition rates in excess of 98%. Similarly, the 3-speaker network achieves a 98.1% recognition rate. While the six speaker TDNN does realize a performance improvement, its recognition rate remains relatively low at 96.5%. We believe that this is due to the high variability of speech data taken from a mixture of male and female speakers.

In addition to the median 23.7% reduction in the number of MSE generated misses using conflict arbitration, we find that the combined classification technique flags 67.7% of the post-arbitration misses as suspect classifications. The cost of flagging these misses is a 7.6% flagging of post-arbitration hits. Figure 6 illustrates these statistics in box-plot form [8] for all the experiments of Table 1. In brief, the box of each plot has vertical extrema that match the first and third quartiles of the sample data; the horizontal line dividing the box delineates the median of the sample data; the inner and (if shown) outer "T"-shaped "fences" of each plot define the outer limits of so-called "adjacent" and "outer" extreme values [8], respectively. Extreme samples falling beyond the outer fence(s) are plotted as dots. The plots show reasonable variance in the sample data, supporting the conclusion that their median values are representative of the conflict-arbitrated MSE/CFM classifier's underlying characteristics.

5 Conclusions

The 98.1% recognition rate for three male speakers represented in a single TDNN is significant because we made no explicit

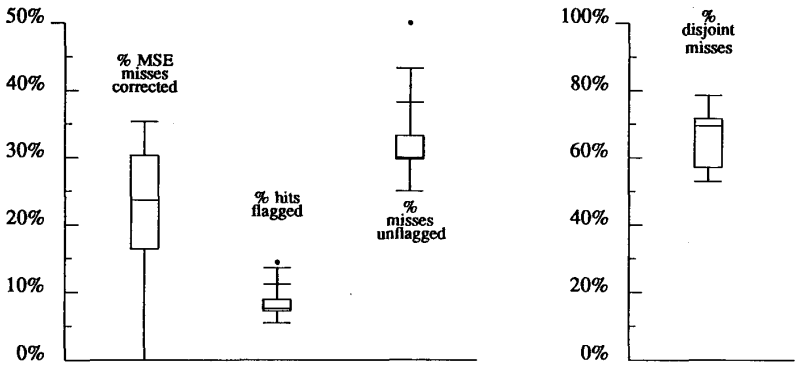

Figure 6: conflict-arbitrated MSE/CFM classification statistics (n=8)

effort to choose three speakers with similar vocal characteristics (beyond choosing 3 males, as opposed to a mix of males and females). This suggests that individual TDNNs might employ conflict arbitration to achieve high recognition rates for specific *classes* of speakers, and that these class-specific networks might then be linked in modular fashion [6,7] to form a speaker-independent network. Clearly there is a limit to the variance of vocal characteristics possible within a given class of speakers. The TDNN trained with six speakers, two of whom were women, displayed recognition performance considerably below that of its 3-speaker counterpart. Figure 6 yields tangential proof of the high variance in the six-speaker data: the box plot depicting the percentage of hits flagged as possible misses has an extreme statistical outlier of 14.4%. This outlier was produced by the six-speaker experiment; it suggests that — independent of classifier type used — the network has a limit to its ability to form generalizations of data with high variance. Indeed, it suggests that there is some trade-off between generalization and key feature extraction — that a balance of both is necessary for robust classification of data characterized by high variance.

We believe that the CFM objective function represents a substantive improvement in connectionist classifier performance. The function is considerably less prone to over-learning than its MSE counterpart. Our initial results also suggest that the CFM classifier is a somewhat faster learner than the MSE classifier (given our choice of parameters). The relative ease with which we have obtained numerically equivalent classification performance with the MSE classifier leads us to believe that there is room for further improvement of the CFM function. We are currently investigating non-sigmoidal mathematical forms in this vein.

Finally, arbitrated classification techniques — that is, classification procedures that evaluate independently-developed, conflicting outcomes, arbitrate a decision, and in the process evaluate their own performance by flagging suspect classifications — represent an effective approach to the complex real-time pattern classification task of speech recognition. These kinds of techniques could form an integral part of large connectionist systems capable of resolving pattern classification ambiguities at many levels of a distributed representation of the speech signal.

6 Acknowledgments

This research was funded in part by grants from Bell Communications Research and ATR Interpreting Telephony Research Laboratories. We wish to thank Dean Pomerleau for numerous insightful comments throughout the course of this research. Additionally, we wish to thank George Gusciora and the WARP group for their tireless support of our computational requirements.

References

[1] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. ASSP-37, March, 1989.

[2] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K., "Phoneme Recognition: Neural Networks vs. Hidden Markov Models," *Proceedings of the 1988 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. S*, pp. 107 - 110, April, 1988.

[3] Rumelhart, D. E., McClelland, J. L., et al., *Parallel Distributed Processing, vol. 1*. Cambridge, MA.: MIT Press,

1987, ch. 8, pp. 322 - 328.

[4] Rumelhart, D. E., Hinton, G. E., and Williams, R. J., "Learning Representations by Backpropagation Errors," *Nature*, vol. 323, pp. 533 - 536, October, 1986.

[5] Ahmad, S. and Tesauro, G., "Scaling and Generalization in Neural Networks: A Case Study," *Proceedings of the 1988 Connectionist Models Summer School*: San Diego, Morgan-Kaufmann Publishers, 1988, pp. 3 - 10.

[6] Waibel, A., Sawai, H., and Shikano, K., "Modularity and Scaling in Large Phonemic Neural Networks", *ATR Interpreting Telephony Research Laboratories technical report TR-I-0034*, August, 1988.

[7] Waibel, A., Sawai, H., and Shikano, K., "Consonant and Phoneme Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks", *Proceedings of the 1989 IEEE International Conference on Acoustics, Speech and Signal Processing, vol. S*, May, 1989.

[8] Tukey, J. W., *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977, ch. 2.