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

JOHN B. HAMPSHIRE, II, STUDENT MEMBER, IEEE, AND ALEXANDER H. WAIBEL, MEMBER, IEEE

Abstract—**This paper presents single and multispeaker recognition results for the voiced-stop consonants /b, d, g/ using time-delay neural networks (TDNN's) with a number of enhancements, including a new objective function for training these networks. The new objective function, which is called the classification figure of merit (CFM), differs markedly from the traditional mean-squared-error (MSE) objective function and the related cross entropy (CE) objective function. Where the MSE and CE objective functions seek to minimize the difference between each output node and its *ideal* activation, the CFM function seeks to maximize the difference between the output activation of the node representing the correct classification and all other nodes (representing incorrect classifications). The results presented here show that each of these three objective functions forms internal representations that differ substantially from those of its counterparts. On the basis of this finding, a simple arbitration mechanism is used with all three objective functions to achieve a median 30% reduction in the number of misclassifications when compared to TDNN's trained with the traditional MSE back-propagation objective function alone. This reduction results in /b, d, g/ error rates that are consistently below 2% for TDNN's trained with individual speakers; it yields a 1.4% error rate for a TDNN trained with three male speakers and a 2.9% error rate for a TDNN trained with six speakers (two female, four male).**

I. Introduction

TIME-delay neural network (TDNN) architectures have been applied to the task of voiced-stop consonant phoneme recognition with excellent results [1]–[3]. In moving from speaker-dependent phoneme recognition to speaker-independent recognition, this paper considers a collection of enhancements to the TDNN that yields improved single speaker and multispeaker 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 back-propagation learning [4], [5]. This new objective function is called the "classification figure of merit" (CFM) in reference to the emphasis it places on the classification result obtained from the network. The mean-squared-error (MSE) and cross entropy (CE) [6] objective functions compare the actual output activations of the network to an *ideal* set of activations for the given input stimulus. They seek to minimize this difference in order to produce the correct output activation corresponding to the correct classification outcome. The CFM objective function, in contrast, uses the ideal output activations only to identify the actual output node corresponding to the correct classification outcome. Once this "correct" node is identified, the CFM function seeks to maximize the difference between it and all of the other (incorrect) nodes.

The results presented here show that the CFM objective function's quantitative performance compares favorably with the MSE and CE classifiers while its qualitative performance is markedly different. Specifically, the different objective functions produce comparable recognition performance, yet they engender substantially different feature abstractions, resulting in largely disjoint misclassified token sets. After training three versions of the same network architecture on the same training set using the three different objective functions, one can use a simple arbitration mechanism to resolve conflicting classification outcomes and reduce by 30% the number of misclassifications made by the MSE classifier alone (unless stated otherwise, all statistics quoted in this paper are median values, owing to the small sample size [$n = 8$]). This mechanism is called "conflict arbitration." The arbitration process identifies or "flags" 80% of the post-arbitration misses, at the cost of flagging 8% of the post-arbitration hits as possible misses (a correct classification is referred to as a "hit," and an incorrect classification is referred to as a "miss"). These enhancements result in single-speaker /b, d, g/ error rates that are consistently below 2%. Additionally, they achieve a 1.4% error rate for a TDNN trained with three male speakers and a 2.9% error rate for a TDNN trained with six speakers (two females and four males).

The experimental conditions under which these findings were made are detailed in [1] and [2]; a condensed version of this work was presented in [7]. Japanese speech data was obtained from six professional announcers (two female, four male), sampled at 12 kHz, parsed for the /b, d, g/ phonemes, and Hamming windowed; from this windowed data 256-point DFT's were computed at 5 ms intervals. The DFT's were used to generate 16 Melscale coefficient spectra at 10 ms intervals. These spectra were normalized to produce suitable input levels for the TDNN's. Training tokens for individual speakers were shuffled randomly and interleaved to produce successive

Manuscript received June 13, 1989; revised March 6, 1990. This work was supported by Bell Communications Research, by ATR Interpreting Telephony Research Laboratories, and by the National Science Foundation under Grant EET-8716324. An earlier version of this paper was presented at the International Joint Conference on Neural Networks, Washington, DC, June 18–22, 1989.

J. B. Hampshire, II, is with the Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, PA 15213-3890.

A. H. Waibel is with the School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA 15213-3890.

IEEE Log Number 9035520.

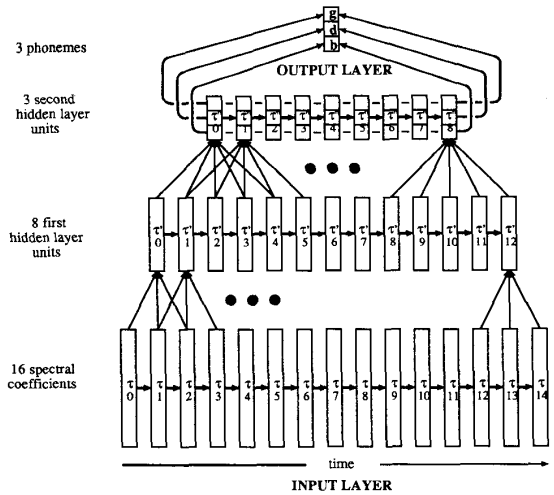

Fig. 1. A time-delay neural network (TDNN) block diagram.

/b, d, g/ tokens (approximately 200 training and 200 testing tokens per phoneme, per speaker). Training tokens for TDNN's trained with multiple speakers were prepared similarly with the additional step of interleaving the tokens across all speakers. Fig. 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], [3].

In this paper we address the following issues. In Section II we review the mathematical forms of the MSE and CE objective functions and discuss why these forms may lead to suboptimal classification performance. In Section III we propose the alternative CFM objective function as a paradigm more suited to the classification task. In Section IV we show experimental results from applying these three objective functions to the task of training TDNN's for single and multispeaker /b, d, g/ phoneme recognition. We show that each objective function yields test data misclassifications that are largely disjoint, and we propose a simple arbitration mechanism by which the number of misclassifications made by a TDNN trained on the MSE objective function alone can be reduced substantially. In Section V we discuss the findings of Section IV and variations on the CFM objective function that might lead to improvements in its classification performance and learning speed. We conclude with a brief summary of our findings.

II. A Review of the MSE and CE Objective Functions

In presenting the CFM objective function, we first review the traditional MSE Objective function used in backpropagation [4], [5] and the closely related CE objective function [6]. The MSE function seeks to minimize the mean squared error between the network's output nodes

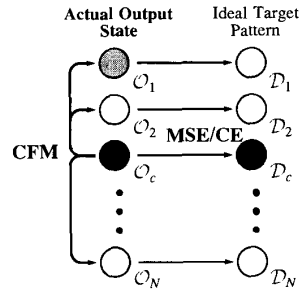

Fig. 2. A graphic comparison of CFM and MSE/CE objective functions.

$\mathcal{O}_1, \cdots, \mathcal{O}_n$ and an ideal or desired set of outputs $\mathcal{D}_1, \cdots, \mathcal{D}_n$ (see Fig. 2)

$$\text{MSE} = 1/N \sum_{n=1}^{N} (\mathcal{O}_n - \mathcal{D}_n)^2 \tag{1}$$

$$\mathcal{E} = 1/2 \sum_{n=1}^{N} (\mathcal{O}_n - \mathcal{D}_n)^2 \tag{2}$$

where (2) is the form used in [4], [5] and $\mathcal{E} = \text{MSE} \cdot N/2$.

The CE objective function views the actual real-valued state of an output node as the probability that the ideal binary output state of the node is "1." It seeks to minimize the difference between actual ($\mathcal{O}_n$) and ideal ($\mathcal{D}_n$) output states by minimizing the cross entropy between the actual and desired probability density functions driving each output node:

$$\text{CE} = -1/N \sum_{n=1}^{N} \{\mathcal{D}_n \log (\mathcal{O}_n) - (1 - \mathcal{D}_n) \log (1 - \mathcal{O}_n)\}. \tag{3}$$

The actual probability density function associated with each output node is of course conditioned by the training set and expressed in the connections of the network. Equation (3) differs from that given in [6] by a constant of $\log (2)/N$. Both of these objective functions engender output states that tend to mimic an ideal pattern. We raise the question of whether the MSE and CE objective functions are optimal for training networks employed as classifiers. We do so on the basis of how their mathematical forms affect the generalizing properties of the networks they are used to train.

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 [8] 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 back-propagation objective function also plays an 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 variance.

For the case of a classifier network with N outputs (representing N possible classes) processing a single input pattern, one can show the following relationship, assuming that the desired output state of the network is binary: for a hit

$$\max \mathrm{MSE}_{\mathrm{hit}} \asymp \frac{N-1}{N} \tag{4}$$

$$\max \mathrm{CE}_{\mathrm{hit}} \asymp \infty \tag{5}$$

and for a miss

$$\min \mathrm{MSE}_{\mathrm{miss}} = 1/(2N) \tag{6}$$

$$\min \mathrm{CE}_{\mathrm{miss}} = \frac{2 \log (2)}{N}. \tag{7}$$

For the CE expression of [6], equation (7) simplifies to $\min \mathrm{CE}_{\mathrm{miss}} = 2$. For both classifiers these thresholds are overlapping, producing regions in the miss domain of state space that yield more optimal values for the objective function than do some regions in the hit domain of state space. In short, neither of these objective functions is monotonic on the hit-miss continuum of state space for $N > 1$. Table I and Fig. 3 help to clarify the concept of monotonicity. Table I lists the output state for three input tokens of a hypothetical two-output network employed as a classifier. Token 1 elicits an ideal classification output with zero MSE and zero CE. Since the network correctly classifies token 1, the result is a hit. The network output is substantially different for token 2; in fact, the activation of $\mathcal{O}_2$ is higher than that of $\mathcal{O}_1$, so this token is misclassified. The MSE and CE scores for this token are 0.303 and 0.799, respectively. The network output for token 3 is again not ideal, but the activation of $\mathcal{O}_1$ exceeds the activation of $\mathcal{O}_2$, so the token is correctly classified. Note that the MSE and CE scores for token 3 are 0.363 and 0.974, respectively—significantly higher than the scores for token 2. Both objective functions yield scores for token 3 that are worse (higher) than the scores for token 2, despite the fact that token 3 is a hit, while token 2 is a miss. For this reason, the MSE and CE objective functions are nonmonotonic on the continuum connecting the best-case ($\mathcal{O}_1 = 1$, $\mathcal{O}_2 = 0$) and worst-case ($\mathcal{O}_1 = 0$, $\mathcal{O}_2 = 1$) output states of the network for an input belonging to class 1. We call this continuum the "hit-miss continuum" of state space.

Fig. 3 illustrates the outputs $\mathcal{O}_1$ and $\mathcal{O}_2$ for tokens 1–3

TABLE I
COMPARISON OF THREE STATES OF A TWO-OUTPUT NETWORK EMPLOYED AS A CLASSIFIER: IN EACH CASE, THE INPUT OF THE NETWORK IS FROM CLASS 1. MEAN-SQUARED ERROR AND CROSS ENTROPY ARE SHOWN FOR EACH OUTPUT STATE

Token	Class	$\mathcal{O}_1$	$\mathcal{O}_2$	MSE	CE	Hit/Miss
1	1	1	0	0	0	Hit
2	1	.45	.55	.303	.799	Miss
3	1	.95	.85	.363	.974	Hit

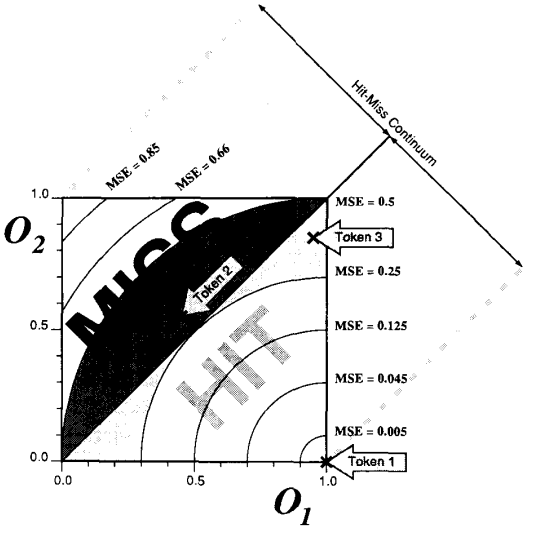

Fig. 3. A contour plot of MSE for a two-output network; output $\mathcal{O}_1$ represents the correct class.

in the network's output state space. Superimposed on this state space are contours of constant MSE for the network when the true class of the network input is class 1 (i.e., when the ideal output state for the network is $\mathcal{O}_1 = \mathcal{D}_1 = 1$, $\mathcal{O}_2 = \mathcal{D}_2 = 0$). The miss domain of state space is therefore the region for which $\mathcal{O}_1 < \mathcal{O}_2$, while the hit domain is the region for which $\mathcal{O}_1 > \mathcal{O}_2$. The dark shaded region of "miss space" delineates the fraction of miss space with lower MSE than some portion of the light shaded region of "hit space." That is, for every point in the light shaded region of hit space, there is some fraction of miss space that yields lower MSE. The shaded regions of Fig. 3 are shaped somewhat differently for the CE objective function; indeed, a higher percentage of state space is shaded for the CE objective function, so it is less monotonic than the MSE objective function. For networks with more than two outputs, the boundaries of the shaded regions in Fig. 3 becomes (hyper)spherical for the MSE objective function. As the number of outputs N in the classifier network becomes large, the fraction of miss space that is shaded increases. In the limit, for every point in hit space there exists some subregion of miss space with lower MSE. Again, the shaded regions of state space for the CE objective function have convex aspherical boundaries, so networks trained with this objective function ap-

proach their nonmonotonic limit, as N grows, more rapidly than their MSE-trained counterparts.

Table I and Fig. 3 indicate that it is possible, at least in principle, for a network trained on a representative training set drawn from an infinitely large ensemble with high variance to minimize its MSE or CE objective function for all training tokens without minimizing the number of misclassifications it makes on a disjoint test set [9]. In Section V we discuss possible characteristics of the random process being classified that would give rise to the nonmonotonic behavior described above. In that section we cite experimental evidence to suggest that nonmonotonic behavior is manifest in some cases of "overlearning" (i.e., when network recognition performance on a disjoint test set peaks and then degrades, while training set performance continues to improve).

One typically attributes poor generalization to a training set that is not truly representative of the ensemble from which it is drawn. Hence, one attempts to improve generalization in one of three ways: 1) by expanding the training set, using its statistical variance to obscure "idiosyncratic" features (i.e., those that are not representative of the ensemble); 2) by "whitening" the training set with noise that obscures idiosyncratic features [10]; or 3) by explictly selecting the training set, choosing tokens that are most representative of the ensemble and most effective in developing optimal classification boundaries [8]. The first solution is rarely possible because one does not have access to a sufficiently large sample set, and the second solution requires careful adjustment of the variance of the noise source in order to eliminate idiosyncrasies without obscuring truly representative features. The third solution requires *a priori* knowledge of those features that are representative of the ensemble; this presents a paradox, since one is often 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 that the mathematical form of the objective function plays an important role in forming general representations of training data for networks employed as classifiers, and that a key element of its mathematical form is the degree to which the function is monotonic on the N-dimensional hit-miss continuum.

III. The CFM Objective Function

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

1) 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 (Fig. 2, $\mathcal{O}_c$) has 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.

2) 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.

3) 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 (Fig. 2, left bank of nodes). 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:

$$\mathrm{CFM} = \frac{1}{N-1} \cdot \sum_{\substack{n=1 \\ n \neq c}}^{N} \alpha\left[1 + \exp\left(-\beta\Delta_n + \zeta\right)\right]^{-1} \quad (8)$$

where

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

Thus

$$\frac{\partial \mathrm{CFM}}{\partial \mathcal{O}_n} = \frac{-\alpha\beta}{N-1} \cdot y_n(1 - y_N) \quad (9)$$

$$\frac{\partial \mathrm{CFM}}{\partial \mathcal{O}_c} = \frac{\alpha\beta}{N-1} \sum_{\substack{n=1 \\ n \neq c}}^{N} y_n(1 - y_N) \quad (10)$$

where

$$y_n \equiv \left[1 + \exp\left(-\beta\Delta_n + \zeta\right)\right]^{-1}.$$

Equations (8) through (10) are variants of the well-known sigmoid function and its derivative [4], [5]. Fig. 4 illustrates the CFM function over the $[-1, 1]$ domain of Δ_n for some representative parameter values. Clearly there are many other functions that meet the CFM specifications itemized above; we present this particular form as an archetype from which further developments might be made. As mentioned earlier, (8) through (10) form a mathematically continuous expression of the degree to which the 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 with some "ideal" output state. From (8) one can show that for a hit

$$\min \mathrm{CFM}_{\mathrm{hit}} \asymp \mathrm{CFM}_n(0) \quad (11)$$

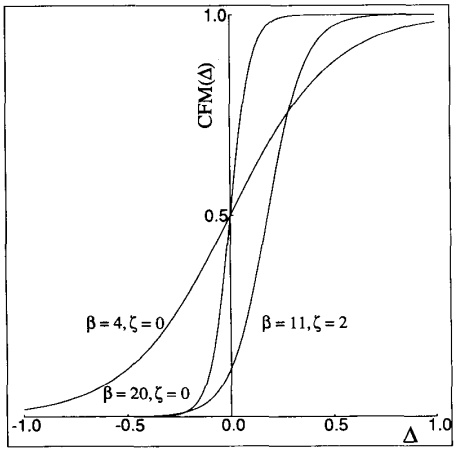

Fig. 4. CFM plotted for representative parameter values: $\alpha = 1.0$, $\beta = 4.0$, $\zeta = 0.0$.

and for a miss

$$\max \mathrm{CFM}_{\mathrm{miss}} = \frac{1}{N-1}\left\{(N-2) \cdot \mathrm{CFM}_n(1) + \mathrm{CFM}_n(0)\right\} \quad (12)$$

where

$$\mathrm{CFM}_n(\Delta) \triangleq \alpha\left[1 + \exp\left(-\beta\Delta_n + \zeta\right)\right]^{-1} \quad (13)$$

a single term from the sum of (8) (the CFM paradigm requires that $N \geq 2$). Table II reviews the output states for the hypothetical two-output network described in Table I of Section II. In Table II, scores are shown for the CFM objective function in (8) with parameters $\alpha = 1.0$, $\beta = 4.0$, $\zeta = 0.0$. Increasing CFM scores indicate an increasingly ideal output state. Clearly, the CFM objective function yields a score that is monotonic on the hit miss continuum. Equations (11) and (12) confirm that the CFM objective function is monotonic on the hit-miss continuum for $N = 2$ (see Fig. 5). For $N \geq 3$, the CFM objective function of (8) is nonmonotonic, but a straightforward modification to its form, described in Section V, produces a CFM variant that is monotonic for networks of arbitrary output dimensionality.

Because one seeks to maximize the CFM objective function, the weight-deflection equation of [4], [5] must be changed to perform gradient ascent (as opposed to gradient descent):

$$\Delta\mathcal{W}(t) = +\epsilon\frac{\partial \mathrm{CFM}}{\partial\mathcal{W}(t)} + \alpha\Delta\mathcal{W}(t-1). \quad (14)$$

The parameter β of (8) 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 decreasing marginal credit it assigns to an increasingly strong hit as well as the amount of decreasing marginal 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). The α parameter is a simple scaling factor, typically equal to unity. Our initial results show that the CFM classifier is quite responsive to changes in the β parameter. In particular, large values of β (≈ 20) engender connection strengths that yield remarkedly weak marginal hits with high MSE in both training and testing data, while smaller values of β (≈ 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. Furthermore, since large values of β yield an increasingly discontinuous CFM function, they tend to engender slow, unstable searches. Although the detailed effects of altering ζ are not well known, we include it in (8)–(10) 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

TABLE II
COMPARISON OF THREE STATES OF A TWO-OUTPUT NETWORK EMPLOYED AS A CLASSIFIER: IN EACH CASE, THE INPUT OF THE NETWORK IS FROM CLASS 1. CFM IS SHOWN FOR EACH OUTPUT STATE

Token	Class	O_1	O_2	CFM	Hit/Miss
1	1	1	0	.982	Hit
2	1	.45	.55	.401	Miss
3	1	.95	.85	.599	Hit

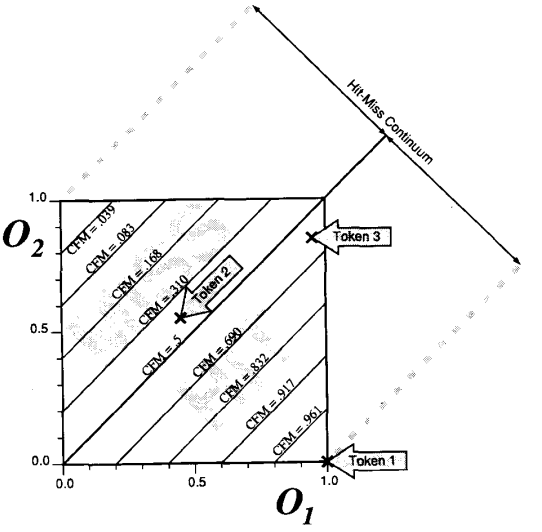

Fig. 5. A contour plot of CFM ($\alpha = 1.0$, $\beta = 4.0$, $\zeta = 0.0$) for a two-output network; output O_1 represents the correct class.

$\alpha = 1.0$ yield recognition rates for the /b, d, g/ recognition task that compare the favorably to those for the MSE and CE classifiers (see Section IV-A). These parameter choices effectively reduce the CFM to a one-parameter function. For $4 \leq \beta \leq 20$ we have not yet found evidence to suggest that the CFM function exhibits the overlearning tendency of the MSE and CE functions—a definite advantage for the CFM function which we discuss further in Section V.

IV. Experimental Results

A. Objective Function Comparisons

Table III shows the results of training a TDNN with tokens from six individual speakers as well as two combinations of speakers using the MSE, CE, and CFM objective functions. Error rates for MSE and CE-trained TDNN's are based on training sessions monitored for the inception of overlearning (i.e., training was monitored for optimal recognition performance on the disjoint test set in order to discount the effect of overlearning). CFM recognition rates are based on unmonitored training sessions. Under these conditions, we find the error rates for the three classifiers roughly equivalent. The median single-speaker CFM error rate is lower than those of the other two functions, and the total number of errors (summed over all six single-speaker trials) for the CFM function is 14% and 19% lower than it is for the MSE and CE functions, respectively. Fig. 6 displays the single-speaker error statistics of Table III in box plot form [11]. 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 [11], respectively. Extreme samples falling beyond the outer fence(s) are plotted as dots. Table III and Fig. 6 suggest that the CFM objective function yields higher classification performance. A one-sided paired t-test [12], [13] of the hypothesis that the mean CFM error rate is significantly lower than those of its competitors is rejected for $p = 0.95$ but accepted for $p = 0.90$. A two-sided paired t-test fails to reject the hypothesis that the mean MSE and CE error rates are identical for $p > 0.70$. In summary, we find that the CFM classifier compares favorably with its counterparts, particularly because we monitored the MSE and CE training sessions for optimal test set performance. The multispeaker results in Table III for the three different classifiers are virtually identical; we discuss possible reasons for this in Section V.

In developing the CFM classifier, our principal 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 support this assertion. First, we take a number of weight vectors for fully MSE-trained TDNN's and use these as input weight vectors for CFM training sessions. The CFM classifier consistently evaluates these initial weight vectors as suboptimal, yielding final CFM weight vectors that are substantially different from their MSE starting points. It is not unusual to find CFM weight vectors computed in this manner nearly orthogonal to their initial MSE values. Additionally, we consistently find that the set of MSE misses and the set of CFM misses are largely disjoint.

The scatter plots of Figs. 7 and 8 illustrate this phenomenon for all phonemes (b, d, and g) of the TDNN trained with three speakers, listed in Table III. The results for this network are representative of the other trials in Table III. Each plot shows the level of activation for the most active bogus output node (i.e., the most active node that does not represent the correct classification) versus the level of activation for the correct output node. Thus, the layout of the scatter plots is very similar to that of Fig. 3. 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 "•". Fig. 7 shows results for the MSE classifier and identifies those MSE misses that are CFM hits (○). Likewise, Fig. 8 identifies CFM misses that are MSE hits (□). It is clear from both figures that the two classifiers have few common

TABLE III
Comparison of /b, d, g/ Error Rates for TDNN Trained with MSE, CE, and CFM Objective Functions (CFM Parameters: $\alpha = 1.0$, $\beta = 4.0$, $\zeta = 0.0$)

Network	Speaker	MSE	CE	CFM
TDNN	MAU	1.7	2.0	1.1
	MHT	0.3	0.6	0.5
single	MNM	2.6	2.3	2.8
speaker	FKN	2.4	2.4	2.2
	FSU	1.8	1.9	1.5
	MMS	2.3	2.5	1.5
	median	2.1	2.2	1.5
	total errors	71	75	61
TDNN				
multi	1st 3	2.7	2.9	2.5
speaker	all 6	4.1	4.3	4.1

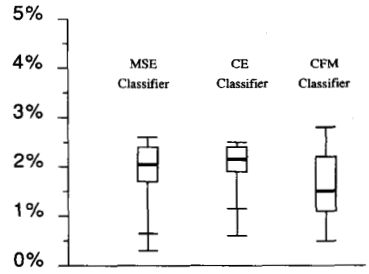

Fig. 6. A comparison of MSE, CE, and CFM /b, d, g/ single-speaker error rates ($n = 6$).

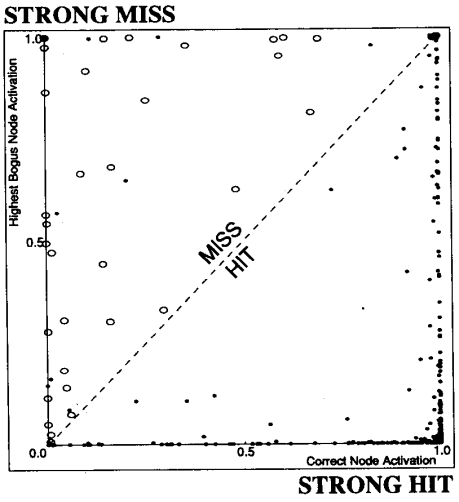

Fig. 7. Scatter plot of MSE classifier outcomes. ○ indicates MSE miss correctly classified by CFM.

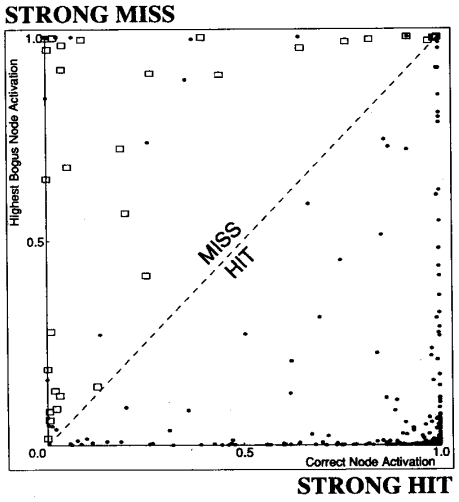

Fig. 8. Scatter plot of CFM classifier outcomes. □ indicates CFM miss correctly classified by MSE.

misses. 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. In fact, the missed token sets of all three classifiers are largely disjoint, as illustrated in the pair-wise comparisons of Fig. 9.

B. *Conflict Arbitration*

One can exploit the disjoint nature of these missed token sets to decrease the number of misclassifications made by the MSE-trained network alone. In [7] we outlined a simple rule-based approach to arbitrating the classification decision when MSE and CFM-trained networks used for recognition yielded conflicting classification outcomes. Since then we have tried a number of arbitration schemes (some involving arbitration networks); we have

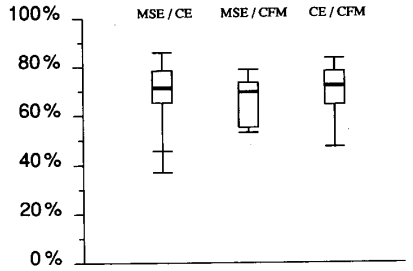

Fig. 9. Percentage of missed test tokens that are disjoint for pair-wise combinations of classifiers ($n = 8$).

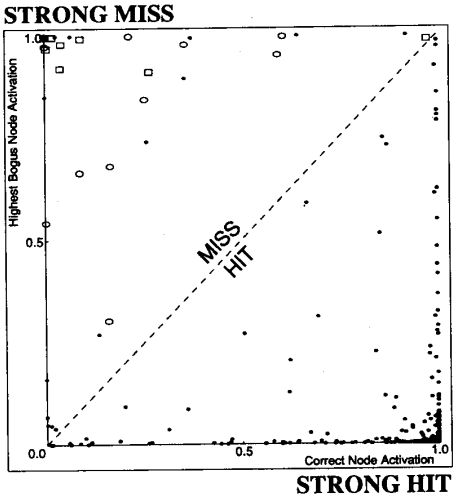

Fig. 10. Scatter plot of arbitrated MSE/CFM classifier outcomes. □ indicates post-arbitration miss correctly classified by MSE. ○ indicates post-arbitration miss correctly classified by CFM.

found that the most effective and most computationally efficient means of arbitration is simply the summation of the two classifiers' output states. Fig. 10 illustrates the reduction in MSE misclassifications (Fig. 7) achieved through this form of conflict arbitration using the MSE and CFM objective functions. Comparing Fig. 10 with Figs. 7 and 8, one can see that the arbitration scheme is particularly effective in eliminating those misses that fall along the hit-miss borderline.

Using the summation form of conflict arbitration with all three classifiers (i.e., training three identical TDNN's with identical training data using the three different objective functions, and processing test data with all three networks, summing their outputs for the final classification decision), we correct a median 30% of the misses made by the MSE classifier alone. Fig. 11 illustrates the 3-way arbitration implementation. The MSE, CE, and CFM-trained TDNN's all share the same input layer and develop independent outputs (representing the phoneme classification outcome) which are then summed and normalized (i.e., divided by the number of classifiers—3 in the case of the 3-way arbitration scheme illustrated). In Fig. 11 the ambiguous MSE classification is corrected by

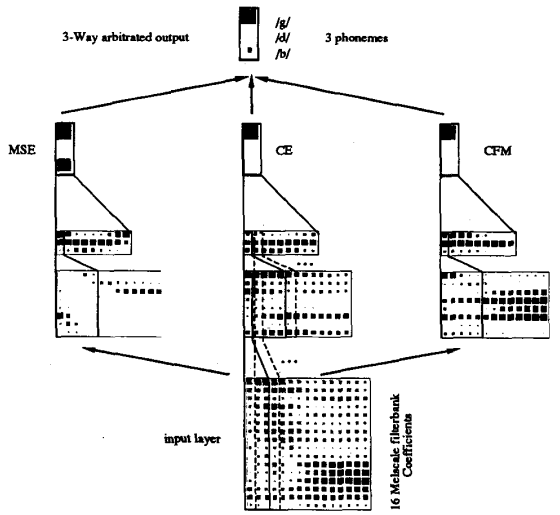

Fig. 11. Three-way conflict arbitration using TDNN's trained with the MSE, CE, and CFM objective functions. Note that the ambiguous MSE output is corrected by the arbitration scheme.

arbitration with the unambiguous CE and CFM classifications.

Table IV summarizes the error rates for the MSE classifier, the three possible pair-wise arbitration schemes, and the 3-way arbitration scheme (for the arbitration schemes, right-hand column values show the percentage of MSE errors corrected). At the bottom of the table a summary shows the median error rate and the total number of errors (summed across all single and multispeaker trials) for the MSE classifier and arbitrated classifiers. Fig. 12 graphically compares the error rates of the various classifiers. Three-way arbitration decreases the median error rate from 2.4% to 1.5% by reducing the number of classification errors by 30%. A one-sided paired t-test accepts the hypothesis that the mean error rates of all the arbitrated classifiers is significantly lower than that of the MSE classifier for $p = 0.975$. In fact, this hypothesis is accepted with $p = 0.99$ for all but the MSE/CE arbitrated classifier. Comparing 3-way arbitration with the three 2-way schemes, we find that 3-way arbitration yields a significantly lower error rate than MSE/CFM and MSE/CE arbitration schemes ($p = 0.95$). The comparison of mean error rates for 3-way and CE/CFM arbitration is less clear; 3-way arbitration is not judged significantly better than CE/CFM for $p = 0.95$, but it is judged better for $p = 0.90$. Our empirical results indicate that 3-way arbitration is convincingly superior to all of the 2-way arbitration schemes for multispeaker tasks. We attribute this to the higher acoustic variance of multispeaker data, which generates a higher proportion of ambiguous classification outcomes for the individual classifiers; 3-way arbitration proves more effective at resolving these ambiguities than 2-way arbitration.

For all the arbitration schemes in Table IV we have found that comparing the assumed post-arbitration min Δ_n of (8) to a threshold provides an effective means of flagging tokens that are possible post-arbitration misses. Since one does not know *a priori* which output node represents the correct classification when one is processing test data, one assumes that the network has yielded the correct classification for the purpose of this differential comparison. If min Δ_n falls below the threshold, the token is flagged as a possible miss. This scheme consistently flags 80% of the post-arbitration misses, at the cost of flagging 8% of the post-arbitration hits as possible misses.

TABLE IV
A SUMMARY OF MSE VERSUS CONFLICT ARBITRATION RESULTS FOR THE MSE, CE, AND CFM CLASSIFIER COMBINATIONS (‖ERROR RATE|PERCENT MSE ERRORS CORRECTED‖)

Network	Speaker	MSE	MSE/CFM		MSE/CE		CE/CFM		3-WAY	
TDNN	MAU	1.7	1.2	27.3	1.8	-9.1	1.5	9.1	1.4	18.2
	MHT	0.3	0.3	0.0	0.5	-50.0	0.2	50.0	0.3	0.0
single	MNM	2.6	2.2	17.6	2.0	23.5	2.2	17.6	1.5	41.2
speaker	FKN	2.4	1.7	26.7	1.9	20.0	1.7	26.7	1.4	40.0
	FSU	1.8	1.6	9.1	1.6	9.1	1.3	27.3	1.5	18.2
	MMS	2.3	1.4	40.0	2.1	6.7	1.7	26.7	1.7	26.7
TDNN										
multi	1st 3	2.7	1.7	35.8	2.0	26.4	1.5	43.4	1.4	47.2
speaker	all 6	4.1	3.5	16.4	3.1	32.9	3.1	25.8	2.9	30.2
SUMMARY										
Median % errors		2.4	1.7		2.0		1.6		1.5	
total errors		283	221		224		203		189	

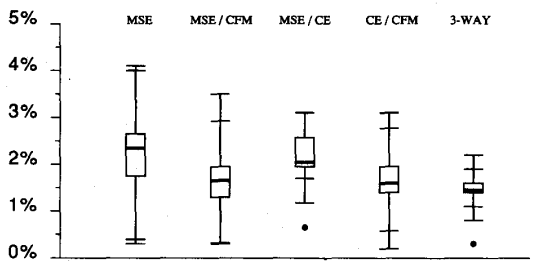

Fig. 12. A graphic comparison of MSE and conflict arbitrated error rates (percent errors, $n = 8$).

C. Control Experiments

The recognition performance improvements afforded by conflict arbitration are significant. However, it is not clear from the error rates alone that the success of arbitration rests on qualitative differences in the internal representations engendered by the different objective functions. It is conceivable that different training sessions using the same objective function but starting at different points in weight space could follow different trajectories through state space to optima that produce disjoint missed test token sets equally effective in conflict arbitration. To test this hypothesis we ran six training sessions on speaker MAU using the MSE objective function. Each of these sessions began at a different, randomly selected point in weight space. Fig. 13 summarizes our findings from using these six trials to generate 15 MSE/MSE arbitrated recognition

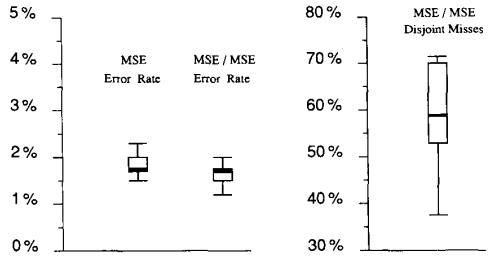

Fig. 13. A graphic summary of an MSE/MSE conflict arbitration control experiment on speaker MAU: error rates and percent disjoint misses (MSE: $n = 6$; MSE/MSE: $n = 15$).

trials on test data. The median percentage of disjoint misses between trial pairs is still quite high (58% versus 70% for arbitration involving different objective functions), but this statistic does not translate to substantive error reductions through arbitration. In fact, the median error rate afforded by MSE/MSE arbitration is 1.7%, identical to that for the MSE classifier alone. Additionally, a one-sided t test fails to reject the hypothesis that the mean error rates for both classifiers are equal for $p \geq 0.90$.

The statistics of Fig. 13 are significant in two ways: the high percentage of disjoint tokens between trials suggests that there are many near-global optima for a given connectionist learning task in weight space; the disparity between the error correction powers of arbitration schemes using different objective functions and one using the same objective function suggests that the efficacy of arbitration rests on fundamental differences in the internal representation generated by different objective functions. It is not enough that the missed token sets used in an arbitration scheme be disjoint, they must also be qualitatively different in order that the postarbitration outcome not be ambiguous.

We have also tested the supposition that the success of arbitration rests on our use of multiple networks to perform the classification task: the argument is that such a method is tantamount to running the classification task on a single network with more hidden units and thus more powers of discrimination. Arbitration does effectively employ a larger network to achieve improved recognition, but it does so by engendering markedly different representations of the training data in what can be thought of as separate subnets. These networks learn independently so that their representations are complementary. This is not the case in a single, larger network. We find that using a single TDNN with twelve units in the first hidden layer instead of eight (see Fig. 1) actually increases error rates by more than 1/2% on any given speaker. Lang and Hinton have obtained similar results and attributed this phenomenon to a TDNN architecture that has too many hidden units, enabling it to learn idiosyncratic details of the training set [3]. This results in degraded powers of generalization and degraded recognition performance on test sets.

V. Discussion

Our initial evaluation of the CFM objective function is encouraging from a number of standpoints. From our trials it appears to perform marginally better on the /b, d, g/ recognition task than the MSE and CE objective functions, particularly for single-speaker trials. We feel that this stems from the issue of the objective function's monotonicity on the N-dimensional hit-miss continuum (i.e., how closely the classification objective and the function used to express that objective match).

A. Possible Conditions Giving Rise to Degraded Classification Performance of Nonmonotonic Objective Functions

In Section II we described the nonmonotonic nature of the MSE and CE objective functions. While their nonmonotonic behavior is unquestionably possible, the conditions that would give rise to such behavior are not so clear. We hypothesize that nonmonotonic behavior can occur when networks are trained to classify stochastic processes. Fig. 14 shows the probabilistic nature of a hypothetical random vector $\boldsymbol{I}$, which is depicted as a scalar for the purposes of illustration. $\boldsymbol{I}$ can belong to one of two classes. Although the prior probabilities of classes 1 and 2 are equal, the class conditional densities $\rho(\boldsymbol{I}|C_1)$ and $\rho(\boldsymbol{I}|C_2)$ are quite different. The class conditional density for class 1 is the lognormal PDF [14] with parameters $\mu_{\log I} = -2.2$, $\sigma^2_{\log I} = 0.7$, while the conditional density for class 2 is the uniform PDF. The differences between the two class conditional densities are extreme for illustrative purposes.

Fig. 14 identifies the optimal (Bayesian) classification boundary ($\boldsymbol{I} = 0.8$) for the random vector $\boldsymbol{I}$ with shaded arrows. Note that the two class conditional densities have some overlap, so the classes are not completely separable. Fig. 15 illustrates a random sample of $\boldsymbol{I}$ (sample size = 2000); the histograms have been normalized so that their areas are unity. One can view these histograms as estimates of the class conditional densities of $\boldsymbol{I}$ in Fig. 14. Note that for this sample size the optimal classification boundary is equivalent to that of Fig. 14. One would expect a network trained on the data depicted in Fig. 15 using the MSE or CE objective functions to form a separating surface at the optimal class boundary, since there are tokens from each class in close proximity to the boundary. We suggest that this may not happen; instead, the network may form a separating surface somewhere to the left of the optimal boundary, within a transition region that contains a relatively minute number of tokens from class 1. This can happen because the size of the network's hidden layer(s) will have been constrained in order to prevent the network from overparameterizing $\boldsymbol{I}$ (i.e., the network architecture will have been restricted in order to prevent poor generalization on disjoint test data). Owing to the large disparity in the numbers of class 1 and class 2 tokens in the vicinity of the optimal classification boundary, the network will minimize its global mean-squared

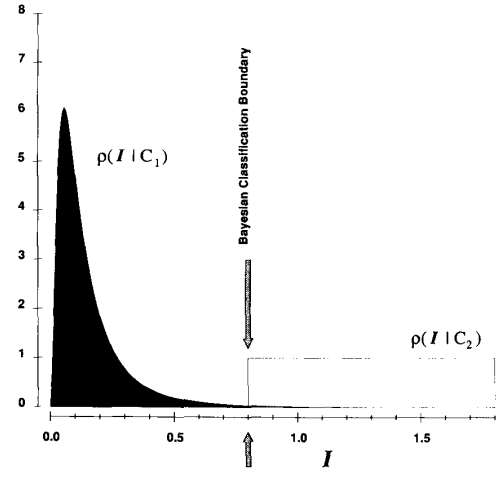

Fig. 14. Class conditional densities for a two-class problem (with equal class "priors").

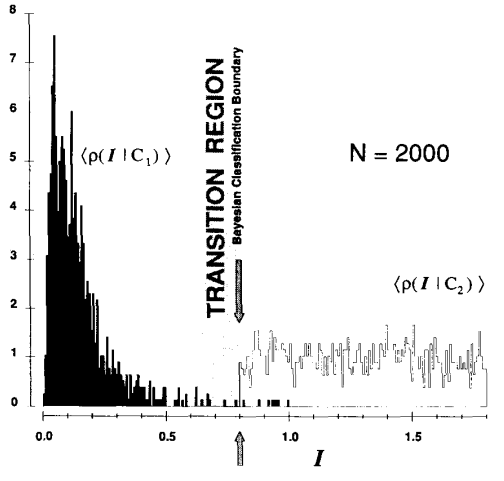

Fig. 15. A sample training set for the two-class problem; 1000 samples are drawn from each class.

error or cross entropy by focusing on the tokens from class 2.

The handful of class 1 tokens in the vicinity of the optimal classification boundary will have their mean-squared error or cross-entropy minimized subject to the class 2 tokens' domination of the error minimization process. This constraint will mean that some test tokens of $\boldsymbol{I}$ in the transition region of Fig. 15 will engender network outputs that fall in the dark shaded region of miss space depicted in Fig. 3—these values of $\boldsymbol{I}$ will produce misses. These misses are not caused by poor network architecture; rather, they are caused by the nonmonotonic nature of the objective function used to train the network. This suggests that overall network classification performance is a function of both optimal network topology and appropriate objective function form. Nonmonotonic behavior may persist as the number of independent tokens used to train the network increases if the network architecture is fixed, particularly if the class-conditional densities of the input random vector are dramatically different (like those in Figs. 14 and 15). However, if the network architecture is augmented as the training set size increases—subject to the constraints imposed by good generalization on test tokens—the network will be able to minimize MSE or CE for the class 1 tokens in the transition region despite the class 2 tokens' domination of the error minimization process. This, in turn, will reduce and ultimately eliminate the misclassifications caused by nonmonotonicity of the objective function (see Section V-B).

We submit that because the CFM objective function is more monotonic on the hit-miss continuum than its MSE and CE counterparts, it avoids the misclassification arising from the phenomenon described above. In Section V-C we describe a variant of the CFM expression in (8) that is strictly monotonic on the hit-miss continuum for networks with arbitrarily large output dimensionality. We are currently conducting a series of experiments designed to directly confirm the hypothesis of this section. Meanwhile, we have indirectly confirmed that the issue of monotonicity plays a role in the over-learning phenomenon. When training with the MSE and CE classifiers, we find that as the global set of network outputs associated with the training token set descends the non-monotonic objective (error) function in state space, most outputs nominally follow a steepest descent trajectory along the error function's surface (e.g., Fig. 3: this is a state-space trajectory, governed by—but not to be confused with—the back-propagation search trajectory in weight-space). However, a small minority of training set outputs and, in effect, a small minority of test set outputs follow trajectories that are not along the objective function's state-space gradient. These trajectories tend to trace a contour within the dark shaded region of miss space depicted in Fig. 3 along which the value of the objective function remains virtually constant. Some of these trajectories appear to stem from borderline crossings from hit to miss domains of state space; in such cases the global error metric is reduced for the vast majority of tokens without any counter-acting increase due to such trajectories. This indicates that the objective function does indeed play a role in the development of general representations, that it has an effect on the degree to which such pathological errors are made. We have not seen these manifestations of overlearning with the CFM function and believe that this is due to the fact that it more closely approximates a truly N-monotonic objective function (see (4)–(7) and (11)–(13)).

B. The Asymptotic Equivalence of Different Objective Functions

In the multispeaker trials the CFM classifier did not perform significantly better than its MSE and CE counterparts. We attribute this to the higher variance of the multispeaker training data obscuring idiosyncratic fea-

tures that, combined with the less monotonic objective functions, would give rise to diminished generalization. Leung and Zue have shown results that support this conclusion [15]. They used a weighted MSE objective function to train a multilayer perceptron for speaker-independent vowel recognition. This weighted MSE objective function was more nearly monotonic on the N-dimensional hit-miss continuum for arbitrary N (a characteristic hereafter referred to as "N-monotonic") than the MSE function. Leung and Zue found that the weighted MSE objective function produced lower error rates for relatively small training sets, but as training set size grew large (corresponding to a large increase in the number of speakers used for training) the error rates for the two classifiers converged.

C. A Monotonic CFM Variant

We have found that the MSE and CE objective functions typically learn to classify training sets perfectly, yet they often produce gross misclassifications (i.e., strong misses) on test data. In contrast, we find that the CFM objective function rarely learns training data perfectly, while it produces a large proportion of test set misses that are predictable by (11) and (12). Indeed, Fig. 8 shows a large number of misses in the upper right-hand corner of the plot: these are borderline misses with CFM $\approx$ max $\mathrm{CFM}_{\mathrm{miss}}$. This indicates that the CFM function is making these borderline errors because the function is nonmonotonic for $N = 3$. We suggest that by virtue of its nonmonotonicity for $N = 3$, the CFM classifier forms classification boundaries on its training set that equate to slightly suboptimal classification boundaries for test data. As a result, a number of test tokens are borderline misses.

One way to assure that the CFM objective function is N-monotonic for arbitrary N is to alter the expression of (8) so that it contains only one term under the summation, the term with the smallest value of Δ_n:

$$\mathrm{CFM}_{N-\mathrm{monotonic}} \triangleq \mathrm{CFM}_n(\min \Delta_n) \tag{15}$$

where $\mathrm{CFM}_n(\Delta)$ is given in (13). We have found that altering the CFM objective function in this manner does reduce test set error rates (converting some of the borderline misses to borderline hits), although the resulting search is considerably slower owing to the reduced magnitude of the gradient of (15) versus that of (8). These improvements are small, on the order of tenths of a percent. In unpublished research [19], K. Lang experimented with what amounts to a linear form of this N-monotonic CFM paradigm. Instead of (15), Lang's objective function was min Δ_n. Lang found that his linear model did not perform as well as the MSE and CE classifiers. However, its proportion of borderline misses to total misses was significantly higher than those of its counterparts (i.e., his N-monotonic classifier made far fewer gross misclassifications than the MSE and CE classifiers did). We hypothesize that the nonlinear CFM form may account for the improved performance results we have seen in our trials.

D. A Faster Learning CFM Variant

The CFM function in (8) learns more slowly than its MSE and CE counterparts. We are currently investigating nonsigmoidal variants of the CFM function to increase its learning speed and its classification accuracy in the nonmonotonic form of (8) and the N-monotonic form of (15). Fig. 16 shows a promising "maximally flat" variant based on the log-magnitude response of the Butterworth filter model [16]. The maximally flat equivalent of (8) is given by

$$\mathrm{CFM}_{\mathrm{MF}} = \frac{1}{N-1} \cdot \sum_{\substack{n=1 \\ n \neq c}}^{N} - \alpha \cdot \log\left\{1 + (\zeta - \Delta_n)^{2\beta}\right\} \tag{16}$$

where all parameters are positive in value and identical in function to those of (8).

Thus

$$\frac{\partial \mathrm{CFM}_{\mathrm{MF}}}{\partial \mathcal{O}_n} = \frac{-2\alpha\beta}{N-1} \cdot \frac{(\zeta - \Delta_n)^{2\beta-1}}{1 + (\zeta - \Delta_n)^{2\beta}} \tag{17}$$

$$\frac{\partial \mathrm{CFM}_{\mathrm{MF}}}{\partial \mathcal{O}_c} = \frac{2\alpha\beta}{N-1} \cdot \sum_{\substack{n=1 \\ n \neq c}}^{N} \frac{(\zeta - \Delta_n)^{2\beta-1}}{1 + (\zeta - \Delta_n)^{2\beta}}. \tag{18}$$

In our inital trials with this maximally flat CFM variant, we find the parameter values of $\alpha = 10$, $\beta = 5$, and $\zeta = 1.5$ (shown in Fig. 16) ($\beta = 5$ denotes a fifth-order Butterworth amplitude response characteristic) yield a maximally flat CFM objective function that learns considerably faster than its sigmoidal counterpart. Table V compares the median MSE, CE, CFM, and $\mathrm{CFM}_{\mathrm{MF}}$ error rates and learning times for six trials of the MAU /b, d, g/ training task (in terms of epochs, or full passes through the training set). The parameter values for the two CFM variants are CFM: $\alpha = 1$, $\beta = 4$, $\zeta = 0$ and $\mathrm{CFM}_{\mathrm{MF}}$: $\alpha = 10$, $\beta = 5$, $\zeta = 1.5$.

These six trials are separate from the single trial presented in Table III—this accounts for the slightly different MAU error rates in Tables III and V. The results of Table V indicate that the maximally flat CFM objective function has a significantly higher error rate than its sigmoidal counterpart. However, this statistic does not reflect the nature of the misses produced by the maximally flat CMF function; more than 75% of these misses are borderline misses. This suggests that an N-monotonic form of (16) (i.e., the single term under the summation of (16), identified by (15)) would yield very low error rates while maintaining an acceptable learning speed.

Beyond the CFM objective function alone, we believe that our results show compelling evidence that the different objective functions engender consistently different internal representations in the networks they are used to train. In this light, conflict arbitration offers a means of avoiding a substantial number of misclassifications that the MSE classifier would make alone. Additionally, it provides a sensitive and relatively specific means of iden-

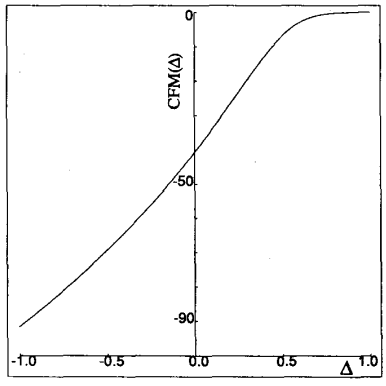

Fig. 16. A nonsigmoidal, "maximally flat" variant of the CFM function ($\alpha = 10, \beta = 5, \zeta = 1.5$).

TABLE V
A COMPARISON OF MEDIAN ERROR RATES AND MEDIAN TRAINING EPOCHS REQUIRED FOR THE MSE, CE, CFM, AND MAXIMALLY FLAT CFM OBJECTIVE FUNCTIONS (SPEECH FROM MAU, $n = 6$)

	MSE	CE	CFM	CFM_{MF}
median error rate	1.7%	1.7%	1.4%	1.8%
median epochs to train	768	526	1186	518

tifying post-arbitration misses—a feature that could prove useful for resolving phoneme recognition ambiguities at a higher level of a hierarchical classification mechanism. We have employed conflict arbitration extensively in a hierarchical connectionist architecture known as the "Meta-Pi" network [17], [18]. This architecture uses multiple conflict-arbitrated TDNN modules to achieve multi-speaker (males and females) phoneme recognition at speaker-dependent recognition rates. These arbitration-based improvements are not without their price, they require that three networks be trained on each training set. Nevertheless, we feel that the benefit warrants the cost, as the 1.4% error rate for three male speakers represented in a "single" arbitrated TDNN illustrates. We made no effort to choose three speakers with similar vocal characteristics (beyond choosing three males, as opposed to a mix of males and females). Clearly there is a limit to the acoustic variance allowable within a given class of speakers. The TDNN trained with six speakers (two of whom were women) displays recognition performance considerably below that of its three-speaker counterpart. This degraded performance corresponds to 15% of the post-arbitration hits being flagged as possible misses, almost twice the percentage of post-arbitration hits flagged for single-speaker trials. This statistic is tangential proof of the high acoustic variance of voiced stop speech from a mix of male and female speakers.

VI. CONCLUSIONS

We believe that the CFM objective function represents a substantive improvement in connectionist classifier performance. The function is less prone to overlearning than its MSE and CE counterparts. Our initial findings suggest that it is also better at a forming general representations of training data. We attribute these characteristics, at least in part, to the fact that it is a closer approximation to a monotonic function on the N-dimensional hit-miss continuum than are the MSE and CE objective functions.

Arbitrated classification techniques—that is, classification procedures that evaluate independently developed outcomes, arbitrate a decision when those outcomes conflict, 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.

ACKNOWLEDGMENT

The authors wish to thank Bell Communications Research, ATR Interpreting Telephony Research Laboratories, and the National Science Foundation for their support of this research. They also wish to thank D. Pomerleau for his numerous insightful comments throughout the course of this research, K. Lang for sharing his extensive research findings with them, and B. Pearlmutter for his incisive and thought-provoking perspectives on a number of issues fundamental to the evolution of this paper. Finally, they wish to thank G. Gusciora, P. Lieu, and the WARP group for their tireless support of the computational requirements.

REFERENCES

[1] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-37, pp. 328–339, Mar. 1989.

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition: Neural networks versus hidden Markov models," in *Proc. 1988 IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. S, Apr. 1988, pp. 107–110.

[3] K. Lang and G. Hinton, "A time-delay neural network architecture for speech recognition," Carnegie-Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-88-152, Dec. 1988.

[4] D. E. Rumelhart *et al.*, *Parallel Distributed Processing*, vol. 1. Cambridge, MA.: M.I.T. Press, 1987, ch. 8, pp. 322–328.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagation errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[6] G. E. Hinton, "Connectionist learning procedures," Carnegie-Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-87-115 (version 2), Dec. 1987, p. 14.

[7] J. Hampshire and A. Waibel, "A novel objective function for improved phoneme recognition using time-delay neural networks," in *Proc. 1989 Int. Joint Conf. Neural Networks* (Washington, DC), June 18–22, 1989.

[8] S. Ahmad and G. Tesauro, "Scaling and generalization in neural networks: A case study," in *Proc. 1988 Connectionist Models Summer School*. San Diego, CA: Morgan-Kaufmann, 1988, pp. 3–10.

[9] M. Brady and R. Raghavan, "Gradient descent fails to separate," in *Proc. 1988 Int. Conf. Neural Networks*, vol. I, pp. 649–656.

[10] J. Elman and D. Zipser, "Learning the hidden structure of speech," UCSD Institute Cognitive Sciences (ICS), Rep. 8701, Feb. 1987, p. 6.

[11] J. W. Tukey, *Exploratory Data Analysis*. Reading MA: Addison-Wesley, 1977, ch. 2.
[12] W. Hines and D. Montgomery, *Probability and Statistics in Engineering and Management Science*. New York: Wiley, 1980, ch. 10.
[13] M. DeGroot, *Probability and Statistics*, 2nd ed. Reading, MA: Addison-Wesley, 1986, ch. 8.
[14] J. Aitcheson and J. A. C. Brown, *The Lognormal Distribution*. London, England: Cambridge University Press, 1966.
[15] H. Leung and V. Zue, "Some phoneme recognition experiments using artificial neural nets," in *Proc. 1988 IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. S, Apr. 1988, pp. 422–425.
[16] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1988, ch. 7.
[17] J. Hampshire and A. Waibel, "The meta-pi network: Building distributed knowledge representations for robust pattern recognition," Carnegie-Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-89-166, Aug. 1989.
[18] J. Hampshire and A. Waibel, "Connectionist architectures for multi-speaker phoneme recognition," in *Advances in Neural Information Processing Systems*, vol. 2, D. Touretzky, Ed. San Diego, CA: Morgan-Kaufmann, 1990.
[19] K. Lang, 1988, unpublished.

*

John B. Hampshire, II, (S'86) was born in Natick, MA, in 1958. He received the B.S.E.E. degree from the U.S. Naval Academy, Annapolis, MD, in 1980, and the M.S. degree in electrical and biomedical engineering sciences from Thayer School of Engineering, Dartmouth College, in 1988. He is currently studying toward the Ph.D. degree in the Department of Electrical and Computer Engineering, Carnegie-Mellon University, where he is researching connectionist approaches to speaker-independent phoneme recognition.

Following commissioning into the Navy, he spent three years at sea and two years as Executive Assistant to a Washington, D.C., based admiral. He resigned from active Naval service in 1896 and returned to school. His general research interests are in pattern recognition, information theory, and connectionist learning theory.

Mr. Hampshire is a student member of the International Neural Network Society.

*

Alexander H. Waibel (S'79–M'80) was born on May 2, 1956 in Heidelberg, West Germany. He received the B.S. degree in 1979 from the Massachusetts Institute of Technology, Cambridge, the M.S. degree in 1980 from the Massachusetts Institute of Technology, and the Ph.D. degree in electrical engineering and computer science in 1986, from Carnegie-Mellon University, Pittsburgh, PA.

From 1980 to 1985, he was a member of the Computer Science Research Staff at Carnegie-Mellon University. In 1986, he joined the faculty of the Computer Science Department as Research Associate and is now a Research Computer Scientist. He holds a joined faculty appointment in the Center for Machine Translation at Carnegie-Mellon. From May 1987 to July 1988, he worked as an Invited Research Scientist at the ATR Interpreting Telephony Research Laboratories in Osaka, Japan, where he continues to maintain joined responsibilities. His current research interests include speech recognition and synthesis, neurocomputing, machine learning, and machine translation. He has written published works and lectured extensively in these areas and has served as a scientific consultant to several corporations worldwide. He has chaired conferences and workshops, served on various program committees, and acted as a referee on numerous journal papers, conference programs, and grant proposals.

Dr. Waibel is a member of the IEEE Acoustics, Speech, and Signal Processing Society, the IEEE Computer Society, the Acoustical Society of America, the Association for Computational Linguistics and the International Neural Network Society. He is also a member of the Technical Committee of the IEEE Signal Processing Society.