Necessary Corrections in Intransitive Likelihood-Ratio Classifiers

Gang Ji *and* **Jeff Bilmes**
SSLI-Lab, Department of Electrical Engineering
University of Washington
Seattle, WA 98195-2500
{*gang,bilmes*}*@ee.washington.edu*

Abstract

In any pattern classification task, errors are introduced because of the difference between the true generative model and the one obtained via model estimation. One approach to this problem uses more training data and more accurate (but often more complicated) models. In previous work we address this problem differently, by trying to compensate (post log-likelihood ratio) for the difference between the true and estimated model scores. This was done by adding a bias term to the log likelihood ratio that was based on an initial pass on the test data, thereby producing an intransitive classifier. In this work, we extend the previous theory by noting that the bias term used before was sufficient but not necessary for perfect correction. We derive weaker (necessary) conditions that still lead to perfect correction, and therefore might be more easily obtainable. We test a number of new schemes on an isolated word automatic speech recognition task. Results show that by using the bias terms calculated this way, the accuracy of classification substantially improves over the baseline and over our previous results.

1 Introduction

Statistical pattern recognition is often based on Bayes decision theory[4], which aims to achieve minimum error rate classification. Given an observation x, for each classification decision, say class c', a loss function $L(c'|c)$ is associated with the decision if the true answer is class c. The goal is to find a class c^* that minimizes the conditional risk $R(c'|x)$

$$c^* = \underset{c'}{\operatorname{argmin}}\, R(c'|x) = \underset{c'}{\operatorname{argmin}} \sum_{c'} L(c'|c) P(c|x), \tag{1}$$

The zero-one loss function is most commonly used

$$L(c'|c) = 1 - \delta_{c'c},$$

which means that the loss for a correct decision is zero, and for any wrong decision the loss is 1, and δ_{ij} is the Kronecker delta function. With this type of loss function, the decision rule becomes

$$c^* = \underset{c}{\operatorname{argmax}}\, P(c|x) = \underset{c}{\operatorname{argmax}}\, P(x|c)P(c), \tag{2}$$

Submitted to Neural □
Information Processing □
Systems (NIPS)-15, 2002

which is based on the posterior probability and is called the minimum-error-rate decision rule.

In multi-class classification, a Bayes' classifier can be seen as a tournament style game, where the winner between "players" is decided using likelihood ratios. Suppose the classes (players) are $\{c_1, c_2, \ldots, c_M\}$, and the observation (game) is x, the winner of each pair of classes is determined by the sign of the log likelihood ratio $L_{ij} = \ln \frac{P(x|c_i)}{P(x|c_j)}$. A practical game strategy can be obtained by fixing an order of comparison, $\{i_1, i_2, \ldots, i_M\}$, where class c_{i_1} plays class c_{i_2}, the winner plays class c_{i_3}, and so on until a final winner is ultimately found. This yields a transitive game, because the the ultimate winner is the same regardless of the comparison order.

To perform the procedure above, however, the correct likelihood ratios are needed, but given a finite amount of training data this is never the case. In previous work [1], we introduced a method to correct for the difference between the true and an approximate log likelihood ratio. In this work, we improve on the method of correction by using an expression that can still lead to perfect correction, but is weaker than what we used before. We show that this condition still achieves a significant improvement over baseline results, on a medium vocabulary isolated word automatic speech recognition task. The paper is organized as follows: Section 2 describes the general scheme and describes past work. Section 3 discusses the weaker correction condition, and its various approximations. Section 4 provides various experimental results on an isolated word speech recognition task. Section 6 provides preliminary results using an iterative update scheme. Finally, Section 7 concludes.

2 Background

A common problem in all machine learning settings is lack of sufficient training — without this, the estimated distribution does not well match the one encountered during testing. The problem is no less acute in speech recognition. In a generative setting, this occurs since instead of the real class conditional model $P(x|c)$, only the estimated quantity $\hat{P}(x|c)$ is available. In the likelihood ratio scheme described above, the log-likelihood ratio that is available for decision making is $\hat{L}_{ij} = \ln \frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)}$ rather than the true log ratio $L_{ij} = \ln \frac{P(x|c_i)}{P(x|c_j)}$

A difference between L_{ij} and $\hat{L}_{ij}$ can cause errors because of inaccuracy in the estimation procedure. One approach to correct for this inaccuracy is to use more complicated class conditional likelihoods, more complicated functional forms of L_{ij}, and/or more training data. In previous work [1], we proposed another approach that requires no change in generative models, no increase in free parameters, no additional training data but still yields improved accuracy. The key idea is to compensate for the difference between L_{ij} and $\hat{L}_{ij}$ using a bias term $\alpha_{ij}(x)$ based on test data[1]:

$$\ln \frac{P(x|c_i)}{P(x|c_j)} = \ln \frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)} + \alpha_{ij}(x) \tag{3}$$

If it is assumed that a single bias term is used for all data, so that $\alpha_{ij}(x) = \alpha_{ij}$, we found that the best α_{ij} is as follows:

$$\alpha_{ij} = \frac{1}{2}\left(D(i\|j) - D(i\|j)\right) - \frac{1}{2}\left(\hat{D}(i\|j) - \hat{D}(i\|j)\right), \tag{4}$$

[1] In this and subsequent analysis, we assume uniform and therefore ignore priors $P(c)$ for simplicity, but these can easily be used if desired.

where $D(i\|j)$ is the Kullback-Leibler (KL) divergence[3]. Under the assumption of symmetric KL-divergence for the true model (e.g., equal covariance matrices in the Gaussian case), the bias term can be solved explicitly as

$$\alpha_{ij} = -\frac{1}{2}\left(\hat{D}(i\|j) - \hat{D}(i\|j)\right). \tag{5}$$

We saw how the augmented estimated likelihood ratio leads to an intransitive game [7, 5], and we investigated a number of intransitive game playing strategies. Moreover, we observed that if the correction was optimal, the true likelihood ratios would be obtained which clearly are not intransitive. We therefore hypothesized and experimentally verified that the existence of intransitivity was a good indicator of the occurrence of a classification error.

This general approach can be improved upon in several ways. First, better intransitive strategies can be developed (for detecting, tolerating, and utilizing the intransitivity of a classifier); second, the assumption of symmetric KL-divergence should be relaxed; and third, the above criterion is stricter than required to obtain perfect correction. In this work, we advance on the latter two of the above three possible avenues for improvement.

3 Improved Intransitive Scheme

An $\alpha_{ij}(x)$ that solves Equation 3 is a sufficient condition for a perfect correction of the estimated likelihood ratio since given such a quantity, the true likelihood ratio would be attainable. This condition, however, is stricter than necessary because it is only the sign of the likelihood ratio that is needed to decide the winning class. We therefore should ask for a condition that corrects only for the discrepancy in sign between the true and estimated ratio, as in the following:

$$\text{sgn}\left[\ln\frac{P(x|c_i)}{P(x|c_j)} - \alpha_{ij}(x)\right] = \text{sgn}\ln\frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)}. \tag{6}$$

As can be seen, this condition is weaker than Equation 3, weaker in that any solution to Equation 3 solves Equation 6 but not vice versa. Note also that Equation 6 provides *necessary* conditions for an additive correction term to achieve perfect correction, since any such correction must achieve parity in the sign. Therefore, it might be easier to find a better correction term since the criterion function (and therefore, set of possible α values) is less constrained. As will be seen, however, analysis of this weaker condition is more difficult. In the following sections, we introduce several approximations to this condition.

3.1 The Sign function and its approximations

The main problem in trying to solve for $\alpha_{ij}(x)$ in Equation 6 is the existence of the sign function. In this section, therefore, we work toward obtaining an analytically tractable approximation. Let us at first assume that the function "sgn()" is the standard (+1/-1)-valued sign function (-1 when its argument is less than zero, and +1 otherwise). We obtain an approximation via a Taylor expansion as follows:

$$\text{sgn}(z+\epsilon) = \text{sgn}(z) + \epsilon\text{sgn}'(z) + o(\epsilon) = \text{sgn}(z) + \epsilon(2H(z)-1)' + o(\epsilon) \tag{7}$$

where

$$H(z) = \begin{cases} 1 & \text{if } z > 0, \\ 1/2 & \text{if } z = 0, \\ 0 & \text{if } z < 0, \end{cases}$$

is the unit step function, and $H'(z) = \delta(z)$, where $\delta(z)$ is the Dirac delta function and satisfying $\int f(z)\delta(z)\,dz = f(0)$. Therefore we say that,

$$\text{sgn}(z+\epsilon) = \text{sgn}(z) + 2\epsilon\delta(z) + o(\epsilon)$$

Of course, the Taylor expansion is valid only when the function is continuous and differentiable, otherwise the error terms can be arbitrarily large. If we find and use a suitable continuous and differentiable approximation rather than the discrete sgn function, the above expansion becomes more appropriate. There exists a trade-off, however, between the quality of the sign function approximation (a better sign function should yield a better approximation in Equation 6) and the error cased by the $o(\epsilon)$ term in Equation 7 (a better sign function approximation will have a greater error when higher-order terms are dropped). We therefore expect that there will exist an optimal balance between the two.

Applying this to the left side of Equation 6, we get

$$\operatorname{sgn}\left[\ln\frac{P(x|c_i)}{P(x|c_j)}-\alpha_{ij}\right]\approx\operatorname{sgn}\ln\frac{P(x|c_i)}{P(x|c_j)}-2\alpha_{ij}\delta\left(\ln\frac{P(x|c_i)}{P(x|c_j)}\right).$$

where we have also made the simplifying assumption that $\alpha_{ij}(x)=\alpha_{ij}$ as we did before. This can be plugged into Equation 6, and to find α_{ij}, the expected value of both sides is taken with respect to distribution $P_{ij}(x)$. If the true class of x was c_i, we should use $P_{ij}(x)=P(x|c_i)$ (similarly for c_j). If neither c_i nor c_j is the true class, it does not matter which distribution is used since when used in a game playing strategy, either winner will ultimately play against the true class. Unfortunately, we do not know the true class at any given point, so we integrate with respect to the the average, i.e., $P_{ij}(x)\overset{\Delta}{=}\frac{1}{2}(P(x|c_i)+P(x|c_j))$. This is accurate because 1) again, when the true class is neither c_i nor c_j the distribution used does not influence the result, and 2) in the two class setting (e.g., when we know that either c_i and c_j is the true class), this yields the true distribution $P(x)$ in the case of equal priors. With these assumption, we get:

$$\int\left(\operatorname{sgn}\ln\frac{P(x|c_i)}{P(x|c_j)}-2\alpha_{ij}\delta\left(\ln\frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)}\right)\right)P_{ij}(x)\,dx=\int\operatorname{sgn}\ln\frac{P(x|c_i)}{P(x|c_j)}P_{ij}(x)\,dx-2\alpha_{ij}$$

since the right term on the left of the equality simplifies for this choice of $P_{ij}(x)$. From this, an expression for α_{ij} can be obtained, as:

$$\alpha_{ij}=\frac{1}{2}\int\operatorname{sgn}\ln\frac{P(x|c_i)}{P(x|c_j)}P_{ij}(x)dx-\frac{1}{2}\int\operatorname{sgn}\ln\frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)}P_{ij}(x)dx$$

The first term is quite similar to the term we saw in the KL-divergence case, shown in the first part of Equation 4. Again, because we have no information about the true class conditional models, we assume this term to be zero (denote this as assumption B). Comparing this assumption with the corresponding one for the KL-divergence case (assumption A), we can see that in general they are not identical. In the Gaussian case, however, we can show that A implies B.

Suppose the models are Gaussian so that $P(x|c_i)=N(\mu_i,\sigma_i^2)$. As shown in [1], assumption A becomes:

$$\frac{\sigma_i^2}{\sigma_j^2}-\frac{\sigma_j^2}{\sigma_i^2}+(\mu_i-\mu_j)^2\left(\frac{1}{\sigma_j^2}-\frac{1}{\sigma_i^2}\right)=0$$

which is true if and only if the variances are identical. Now assume assumption B, that the covariances are identical, and (without loss of generality) that $\mu_i<\mu_j$. The point $\bar{x}$ where $P(\bar{x}|c_i)=P(\bar{x}|c_j)$ is $\bar{x}=\frac{1}{2}(\mu_i+\mu_2)$ and hence $P(x|c_i)>P(x|c_j)$ when $x<\bar{x}$ and $P(x|c_i)<P(x|c_j)$ when $x>\bar{x}$. Therefore,

$$\int\operatorname{sgn}\ln\frac{P(x|c_i)}{P(x|c_j)}(P(x|c_i)+P(x|c_j))\,dx=\int_{-\infty}^{\bar{x}}(P(x|c_i)+P(x|c_j))\,dx-\int_{\bar{x}}^{+\infty}(P(x|c_i)+P(x|c_j)).$$

This is zero because $\int_{-\infty}^{\bar{x}} P(x|c_i)\,dx = \int_{\bar{x}}^{+\infty} P(x|c_j)\,dx$ and $\int_{\bar{x}}^{+\infty} P(x|c_i)\,dx = \int_{-\infty}^{\bar{x}} P(x|c_j)\,dx$. Therefore, A implies B.

We have found that the opposite (B implies A) is not true in general, so that assumption B is less restrictive than A, at least in the Gaussian case.

Under this assumption, we can produce an expression for the resulting α_{ij} using the weak law of large numbers as follows:

$$\alpha_{ij} \approx \frac{1}{2}\left(\frac{1}{N_i}\sum_{x\in c_i} \operatorname{sgn}\ln\frac{\hat{P}(x|c_j)}{\hat{P}(x|c_i)} + \frac{1}{N_j}\sum_{x\in c_j} \operatorname{sgn}\ln\frac{\hat{P}(x|c_j)}{\hat{P}(x|c_i)}\right), \tag{8}$$

where N_i and N_j are number of samples from model c_i and c_j respectively. And like in [1], since the true classes are unknown, we perform an initial classification pass to get estimates and use these in Equation 8.

Note that there are two potential sources of error in the analysis above. The first is assumption B, which we argue can be a less severe approximation than in the KL-divergence case. The second is the error due to the discontinuity of the sign function. To address the second problem, rather than using the sign function in Equation 8, we can approximate it a with number of continuous differential functions in the hope of balancing the trade-off that was mentioned above. On the left in Figure 1, we show three sign-function approximations, hyperbolic tangent, arctangent, and a shifted sigmoid function. On the right of the figure, the shifted sigmoid is presented with several values of its free parameter $beta$. In the following, we derive expressions for α_{ij} for each of these sign approximations.

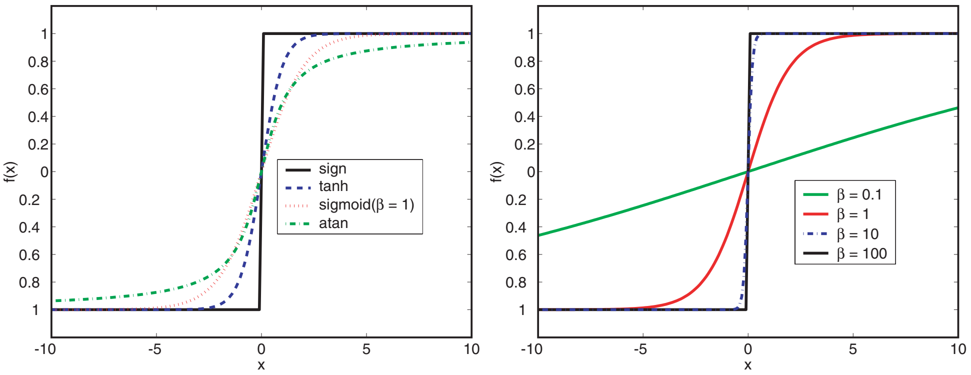

Figure 1: Left: Several approximating sign function. Right: Shifted sigmoid with different β values.

3.2 Hyperbolic tangent

In the hyperbolic tangent case:

$$\operatorname{sgn} z \approx \tanh z = \frac{e^z - e^{-z}}{e^z + e^{-z}}.$$

which yields:

$$\alpha_{ij} \approx= \frac{1}{2}\int \frac{\hat{P}^2(x|c_j) - \hat{P}^2(x|c_i)}{\hat{P}^2(x|c_i) + \hat{P}^2(x|c_j)}(P(x|c_i) + P(x|c_j))\,dx$$

or again using the law of large numbers,

$$\alpha_{ij} \approx \frac{1}{2}\left(\frac{1}{N_i}\sum_{x\in c_i}\frac{\hat{P}^2(x|c_j)-\hat{P}^2(x|c_i)}{\hat{P}^2(x|c_i)+\hat{P}^2(x|c_j)}+\frac{1}{N_j}\sum_{x\in c_j}\frac{\hat{P}^2(x|c_j)-\hat{P}^2(x|c_i)}{\hat{P}^2(x|c_i)+\hat{P}^2(x|c_j)}\right). \qquad (9)$$

3.3 Arctangent

We can also replace the sign function by arctangent.

$$\text{sgn}x \approx \frac{2}{\pi}\tan^{-1}x$$

In this case,

$$\alpha_{ij} \approx \frac{1}{\pi}\left(\frac{1}{N_i}\sum_{i\in c_i}\tan^{-1}\ln\frac{\hat{P}(x|c_j)}{\hat{P}(x|c_i)}+\frac{1}{N_j}\sum_{i\in c_j}\tan^{-1}\ln\frac{\hat{P}(x|c_j)}{\hat{P}(x|c_i)}\right). \qquad (10)$$

3.4 Shifted Sigmoid

Sigmoid function is the solution of the differential equation $y' = y(1-y)$ and has the form $f(x) = \frac{1}{1+e^{-\beta x}}$, where the free parameter β (the inverse temperature) determines how well the curve will approximate a discontinuous function (see Figure 1 right). Using the sigmoid function, we can approximate the sign function as

$$\text{sgn}x \approx \frac{2}{1+e^{-\beta x}} - 1,$$

Note that the approximation improves as β increases. Hence,

$$\alpha_{ij} \approx \frac{1}{2}\left[\frac{1}{N_i}\sum_{i\in c_i}\left(1-\frac{2}{1+\frac{\hat{P}^\beta(x|c_j)}{\hat{P}^\beta(x|c_i)}}\right)+\frac{1}{N_j}\sum_{i\in c_j}\left(1-\frac{2}{1+\frac{\hat{P}^\beta(x|c_j)}{\hat{P}^\beta(x|c_i)}}\right)\right]. \qquad (11)$$

4 Experimental Evaluation

Similar to previous work, we implemented this technique on NYNEX PHONEBOOK[6, 2], which is a medium vocabulary isolated word speech corpus. A Gaussian mixture hidden Markov model (HMM) is used to calculate the probability scores $\hat{P}(x|c_i)$ where in this case x is a matrix of feature values, and c_i is a given spoken word. The HMM models use four hidden states per phone, and 12 Gaussian mixtures per state. This yields approximately 200K free model parameters in total.

In our experiments, the steps are as follows: 1) calculate $\hat{P}(x|c_i)$ using full HMM inference (no Viterbi approximation) for each test case and for each class (word); 2) classify the examples using just the log likelihood ratios $\hat{L}_{ij} = \log\frac{\hat{P}(x|c_i)}{\hat{P}(x|c_j)}$; 3) calculate the bias term using one of the techniques described above; 4) classify again using the new improved likelihood ratio $S_{ij} = \hat{L}_{ij} + \alpha_{ij}$. In this case, since the procedure is no longer transitive, we run 1000 random games (as in [1]) and choose the most frequent winner as the final winner. The results are shown in Table 1.

In the table, the first column gives the size of the vocabulary (number of different classes) in the test data; the first column shows the baseline which is using the $\hat{L}_{ij}$ criteria; the rest are the results using the bias terms. In the sigmoid case, the inverse temperature is set to $\beta = 1$. As we can see from the table, in all cases the new methods yield a significant improvement in accuracy. For each method, the error rate is about the same after the correction.

vocab	baseline $\hat{L}_{ij}$	sign	tanh	atan	sigmoid($\beta = 1$)
75	2.3358	1.7584	1.7584	1.7581	1.7584
150	3.3107	2.8258	2.8382	2.8269	2.8258
300	5.2251	4.7524	4.7492	4.6984	4.7524
600	7.3927	6.6383	6.6109	6.5972	6.6383

Table 1: Word error rates using different approximations to the sign function.

5 Sigmoid with Different Temperatures

The shifted sigmoid function can be fine-tuned to approximate the sign function by changing the value of β as shown in Figure 1. This function is particularly useful since it allows us to investigate the trade off mentioned in Section 3.1. The results are shown in Table 2 for $\beta = \{0.1, 1.0, 10, 100\}$.

vocab	0.1	1.0	10	100	KL-divergence (from [1])
75	1.8228	1.7584	1.5581	1.5708	1.9147
150	2.6502	2.8258	2.6523	2.4664	2.7228
300	4.7448	4.7524	4.2855	3.9502	4.2893
600	6.6581	6.6383	6.0409	5.6980	5.9144

Table 2: Word error rates using shifted sigmoid function with different β values: $\{0.1, 1.0, 10, 100\}$

From the results we can see that the overall performance increases as we increase the inverse of temperature, β. This is because when β increases, the shifted sigmoid curve is a better approximation to the sign function. For $\beta = 100$, the results here show an improvement over the comparable KL-divergence results reported in [1] (shown in the right-most column in the table). We are currently investigating larger β values to determine when the inaccuracies due to the Taylor error term start affecting the results (note, however, that for the 75 word case, it seems this has occurred at $\beta = 100$).

6 Iterative Classification

The scheme above relies on an initial classification run using $\hat{L}_{ij}$ to obtain the estimates of the class identities on the test set. A second run is then used with the likelihood ratio correction factor S_{ij}, producing improved class identity values. It therefore should be possible to iterate this procedure such that further improvements are obtained (see also [1]). We attempted this procedure using the new likelihood ratio results, as shown in Figure 2. Unfortunately, we do not see any improvement, and in certain cases the error results start to diverge. This behavior is different from what was seen using the KL-divergence criterion in [1]. We plan to investigate and analyze these unexpected results further.

7 Conclusion

We extended the basic log likelihood ratio classification by compensating the difference between real model and estimated model. Rather than focusing on the sufficient correction factor as done in previous work, approximate the exact compensation using continuous differentiable approximations of the sign function. Results show that by adding the bias term under the different approximations, the error rates significantly decrease. Of all the functions we used, the shifted sigmoid has the advantage in that it has a free parameter that

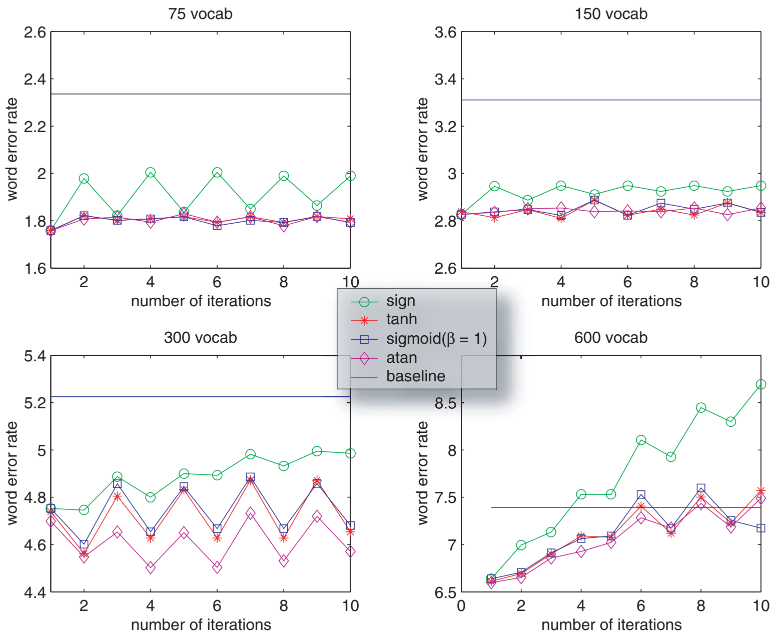

Figure 2: iterative decoding

can be tuned. It is shown that when the parameter β increases, the error rate also decreases. In future work, we will further investigate this function, and will attempt to utilize higher order terms in the Taylor expansion, will apply our methodology on new data sets, and since none of these methods are transitive, we will further analyze why intransitively occurs and how it can potentially be utilized.

References

[1] J. Bilmes, G. Ji, and M. Meilă. Intransitive likelihood-ratio classifiers. In *NIPS*, 14, Vancouver, Canada, December 2001.

[2] J.A. Bilmes. Buried Markov models for speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, March 1999.

[3] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2000.

[5] R. Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey*. Dover, 1957.

[6] J. Pitrelli, C. Fong, S.H. Wong, J.R. Spitz, and H.C. Lueng. PhoneBook: A phonetically-rich isolated-word telephone-speech database. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[7] P.D. Straffin. *Game Theory and Strategy*. The Mathematical Association of America, 1993.