Likelihood-Based Semi-Supervised Model Selection with Applications to Speech Processing

Christopher M. White, *Member, IEEE*, Sanjeev P. Khudanpur, *Senior Member, IEEE*, and Patrick J. Wolfe, *Senior Member, IEEE*

***Abstract*—In conventional supervised pattern recognition tasks, model selection is typically accomplished by minimizing the classification error rate on a set of so-called development data, subject to ground-truth labeling by human experts or some other means. In the context of speech processing systems and other large-scale practical applications, however, such labeled development data are typically costly and difficult to obtain. This article proposes an alternative semi-supervised framework for likelihood-based model selection that leverages unlabeled data by using trained classifiers representing each model to automatically generate putative labels. The errors that result from this automatic labeling are shown to be amenable to results from robust statistics, which in turn provide for minimax-optimal censored likelihood ratio tests that recover the nonparametric sign test as a limiting case. This approach is then validated experimentally using a state-of-the-art automatic speech recognition system to select between candidate word pronunciations using unlabeled speech data that only potentially contain instances of the words under test. Results provide supporting evidence for the utility of this approach, and suggest that it may also find use in other applications of machine learning.**

***Index Terms*—Likelihood ratio tests, pronunciation modeling, robust statistics, semi-supervised learning, sign test, speech recognition, spoken term detection.**

I. Introduction

THIS article develops a simple and powerful likelihood-ratio framework that enables the use of *unlabeled* development data for model selection and system optimization in the context of large-scale speech processing. Within the speech engineering community, *acoustic* likelihoods have long played a prominent role both as a training criterion and an objective function to aid in system development. Log-likelihood ratios have in turn featured ever more prominently in areas such as speech, speaker, and language recognition; for instance, it is now common practice that "target" model likelihoods are compared to those of a universal "background" model as part of many large-scale speech processing systems [1].

A. Model Selection Using Likelihood Ratios

Comparing data likelihoods between competing models can serve as an effective means of model selection for classification and regression tasks. However, when considering conditional likelihoods of the observed data given *labels* such as orthographic transcriptions of speech waveforms, previous work has assumed that orthographic labels have been correctly assigned by human experts, and hence are known exactly. However, such "labeled data" do not come for free; their acquisition requires the time and expertise of a trained linguist, hence limiting scalability to the large sample sizes necessary to succeed in practical speech engineering tasks.

This article thus posits a framework in which likelihoods evaluated using labels that are *automatically assigned* by two competing systems can serve as proxies for likelihoods based on ground-truth labeling. This yields not only a methodologically sound algorithmic framework through which to incorporate unlabeled data into the likelihood-based model selection process, but also practical engineering strategies for selecting between competing models in order to optimize large-scale systems. Experiments to select between candidate word pronunciations in the context of state-of-the-art speech processing systems, using well-known corpora and standard metrics, serve to demonstrate the benefit of unlabeled development data in the context of large-scale speech processing.

To construct this framework, insights from robust statistics are used to formulate the resultant semi-supervised model selection problem in a manner that permits principled analysis, and from which efficient and effective algorithms can be derived. By considering the automatic labeling procedure as a mixture of correct and incorrect assignments, the influence of incorrect labeling can be limited through what is known as a *censored* likelihood ratio evaluation.

The well-known nonparametric sign test arises as a natural limiting procedure in this setting, and the technical development of this article shows how optimality properties derived by Huber [2] can be applied in the semi-supervised setting to ensure that the maximal model selection error induced by automatic labeling is minimized. Thusly one arrives at an algorithmic procedure that compares the relative performance of two competing systems in order to test the significance of performance differences between them, and hence to select the model that is "closest" (in the sense of Kullback-Leibler divergence) to the true data-generating distribution.

B. Unlabeled Data in the Context of Speech Processing

To clarify the notions of supervised/semi-supervised learning and labeled/unlabeled data in the speech processing context at hand, we briefly recall the standard machine learning paradigm as follows. Fundamentally, one assumes

C. M. White is with the Human Language Technology Center of Excellence (HLT-COE), Johns Hopkins University, and the Statistics and Information Sciences Laboratory, Harvard University, Oxford Street, Cambridge, MA 02138 (e-mail: cmwhite@seas.harvard.edu); S. P. Khudanpur is with the Center for Language and Speech Processing, Johns Hopkins University, Charles St., Baltimore, MD 21218 (email: khudanpur@jhu.edu); and P. J. Wolfe is with the Statistics and Information Sciences Laboratory, Harvard University (e-mail: wolfe@stat.harvard.edu). This work was completed while C. M. White was an HLT-COE Graduate Fellow.

the existence of an unknown joint probability distribution $p_{X,Y}(x, y) \neq p_X(x)p_Y(y)$, from which a number of independent and identically distributed samples $(x_1, y_1), (x_2, y_2), \ldots$ are available; these are termed *training data*, and are used to fit a model that predicts values taken by Y based on observed instances of X. In classification tasks Y is a discrete random variable, and its range of possible values comprises the set of *labels*—corresponding to, for example, an orthographic transcript of the word or phrase represented by an instance of acoustic waveform data X.

The goal in a traditional *supervised learning* scenario is to devise algorithms that strike a balance between fidelity to the set of labeled examples $\{(x_i, y_i)\}$, and effective generalization to other as-of-yet unseen *test data* comprising additional observations of X—a classical bias-variance trade-off between model goodness-of-fit and generalization properties. This trade-off is typically optimized by calculating empirical error rates on an additional "held-out" set of labeled data for which ground truth is known, in a manner similar to parameter estimation via cross-validation.

Fitting a model to accomplish this goal is thus mathematically equivalent to building a system, and one speaks of the "training" or model-building stage, and the "testing" or application stage, in which a system is subsequently deployed and put into practical use—and which assumes that both training and test data are drawn from the same probability distribution. When this assumption is satisfied, it is clear that speech engineering systems benefit directly from ever-greater amounts of *labeled training data*. Time, money, and expertise, however, typically limit the amount of such data available in any given application scenario of interest. It is thus of much interest to develop algorithms that are built using some amount of labeled training data, but whose performance can be further improved through careful use of *unlabeled* data—the so-called semi-supervised learning paradigm [3].

Thus far, the application of semi-supervised methods to speech processing has been limited to ideas such as data augmentation [4] or self-training [5], each of which involves re-fitting the models under consideration—and hence rebuilding the corresponding speech engineering systems. While such approaches have shown promise, such extreme re-fitting may not be desirable—or even possible—in certain settings, for instance when a large-scale system is already deployed and must be adapted to new test conditions.

Speech engineering is thus ripe for the introduction of new semi-supervised learning approaches; not only can nearly limitless amounts of acoustic waveform data be acquired from a variety of digital sources, but also many algorithms have matured to the point that performance improvements are often driven simply by increasing the amount of labeled training data. Employing unlabeled data to directly improve existing approaches, however, *requires inferring the labels*—and in this context, a natural but unsolved problem is to understand whether and how *automatically labeled data* taken as output from current systems can be used to this effect. As indicated above, this article brings ideas from robust statistics and likelihood-based model selection to bear on this problem, and introduces not only a framework to analyze the errors resulting from automatic labeling, but also a practical means of treating them.

The article is organized as follows. Section II develops likelihood-based semi-supervised model selection techniques, first considering the case of labeled data, and subsequently the unlabeled case. Section III then formulates this semi-supervised framework in the speech processing context of selecting from amongst competing pronunciation models to optimize system performance. Large-scale experiments with well-known data sets in Section IV then demonstrate that this approach achieves state-of-the-art performance in the context of speech recognition, spoken term detection, and phonemic similarity to a given reference, even when compared to the conventional *supervised* method of forced alignments to reference orthographic transcripts. Section V concludes the article with a discussion of these results and their implication for improving speech processing through the use of unlabeled development data.

II. Theory: Likelihood-Based Model Selection

Viewed from a machine learning perspective, parametric statistical models are directly instantiated as large-scale speech processing systems. Labeled data are used to fit model parameters in the manner described above; e.g., to estimate the state transition matrix of a hidden Markov model. In addition, one must also typically fit a modest number of parameters that alter the structure or function of the model class under consideration; for instance, in automatic speech recognition, the marginal acoustic likelihood of an utterance typically depends on a model for the pronunciation(s) of a given word—a setting we return to in Section III.

When training and test conditions match exactly, all parameters can be fitted simultaneously during the training stage, using principled and efficient procedures such as the expectation-maximization algorithm. In practice, however, it may be the case that only a small amount of labeled training data is well matched to the conditions that prevail during test—precluding even cross-validation as an option—or that a deployed system must be adapted to new test conditions in the absence of its original training data. In such cases it is typical to set aside a small amount of *development data* for purposes of *model selection* as follows.

A. The Supervised Case: Labeled Development Data

Recall that in our setting, X represents acoustic waveform data, and hence is a continuous random variable. The true but unknown data-generating model, then, takes the form of a conditional probability density function $p(x \mid y) := p_{X \mid Y}(x \mid Y = y)$. When interpreted for fixed X as a function of unknown label Y, this density thus evaluates to the acoustic likelihood of X for any given candidate label $Y = y$.

In practice, we have access to $p(x \mid y)$ only through the given pairs of training samples $(x_1, y_1), (x_2, y_2), \ldots$, and we must proceed in the absence of direct knowledge of the true model. Any speech processing system will in turn generate its own set of putative acoustic likelihoods, and thus it is natural to

seek the likelihood function that is closest to the true data-generating model $p(x \mid y)$, in hopes that this will yield the best overall system performance. This leads to a *model selection* problem in which we use the training samples at hand as a proxy for $p(x \mid y)$, to choose amongst competing models and build a system that can predict Y given X with minimal misclassification error.

Assume, then, that we have several competing sets of candidate models $p_1(x \mid y; \theta_1), p_2(x \mid y; \theta_2), \ldots$, each dependent on distinct parameter sets $\theta_1, \theta_2, \ldots$, whose quality we wish to evaluate with respect to the true (but unknown) model $p(x \mid y)$. A natural approach is to evaluate the Kullback-Leibler divergence of the "best" representative $p_k(x \mid y; \theta_k^*)$ of each set from $p(x \mid y)$, with θ_k^* the maximum-likelihood estimate of parameter set θ_k as determined from the training data. Thus we seek

$$\begin{aligned} &\operatorname*{argmin}_k \mathbb{E}_p \left(\log p(x \mid y) \right) - \mathbb{E}_p \left(\log p_k(x \mid y; \theta_k^*) \right) \\ \equiv &\operatorname*{argmax}_k \mathbb{E}_p \left(\log p_k(x \mid y; \theta_k^*) \right), \end{aligned}$$

with $-\mathbb{E}_p \left(\log p_k(x \mid y; \theta_k^*) \right)$ sometimes referred to as the *cross-entropy* of p_k relative to p, and the corresponding optimization task one of *cross-entropy minimization*.

Under the assumption of independent and identically distributed pairs of training examples, we may form an empirical estimate of each cross-entropy simply by evaluating the respective data log-likelihoods $\log p_k(x \mid y; \theta_k^*)$ with respect to each pair of training samples, and forming the corresponding arithmetic averages. Assuming the necessary technical conditions of [6], it then follows that we may formulate a multi-way hypothesis test amongst models $p_1, p_2, \ldots$. We later consider this multi-way setting in detail; however, for clarity of exposition, we first consider the case of only *two* competing models p_1 and p_2, which admits three possible outcomes:

$$\begin{aligned} \mathcal{H}_0 &: \mathbb{E}_p \left(\log p_1(x \mid y; \theta_1^*) \right) = \mathbb{E}_p \left(\log p_2(x \mid y; \theta_2^*) \right) \\ \mathcal{H}_1 &: \mathbb{E}_p \left(\log p_1(x \mid y; \theta_1^*) \right) > \mathbb{E}_p \left(\log p_2(x \mid y; \theta_2^*) \right) \\ \mathcal{H}_2 &: \mathbb{E}_p \left(\log p_1(x \mid y; \theta_1^*) \right) < \mathbb{E}_p \left(\log p_2(x \mid y; \theta_2^*) \right). \end{aligned}$$

Hypothesis $\mathcal{H}_k$ thus favors the kth competing model, with the null hypothesis $\mathcal{H}_0$ representing their equivalence.

The natural test statistic in this *labeled data* setting is then given by the log-ratio of likelihoods p_1, p_2 described above, evaluated with respect to training data—possibly even the same training data used to fit the maximum-likelihood model parameter estimates θ_k^*—as follows:

$$T_{\text{lab}} := \sum_j \log \frac{p_1(x_j \mid y_j; \theta_1^*)}{p_2(x_j \mid y_j; \theta_2^*)}. \tag{1}$$

The careful reader will note that in such a regime, where expectations are defined with respect to some unknown distribution p, we are in fact working with potentially *misspecified* models p_1 and p_2; see [7], [8] for properties of maximum-likelihood estimation of the parameter sets θ_1 and θ_2 in this setting; for our purposes it suffices to note that such estimators still possess the requisite technical properties.

In the case of interest to us here, the conditional models p_1 and p_2 are assumed to be *strictly non-nested*, such that no conditional distribution in X given Y can be achieved by both p_1 and p_2. Vuong [6] shows a central limit theorem for this setting when $\mathcal{H}_0$ is in force, in that as the number of training samples grows large, an appropriately standardized version of the test statistic T_{lab} is asymptotically distributed as a unit Normal. (It is straightforward to proceed in the absence of this assumption, with appropriate adjustments to test statistic asymptotics.) The necessary normalization is given by the sample standard deviation of log-likelihood ratio evaluations times the root of the number of training samples; if $\mathcal{H}_0$ fails to be in force, then the value of this statistic diverges (almost surely) to $\pm\infty$.

This result in turn implies a concrete directional test for model selection: fixing a significance level α yields a corresponding critical value $z_{\alpha/2}$ according to the standard Normal distribution. If the normalized test statistic evaluates to greater than $z_{\alpha/2}$, we select model p_1; if it evaluates to less than $-z_{\alpha/2}$, we decide in favor of model p_2. Otherwise, we conclude that there is insufficient evidence to reject the hypothesis $\mathcal{H}_0$ of model equivalence, and we conclude that models p_1 and p_2 cannot be distinguished on the basis of the given training data and chosen significance level.

B. The Semi-Supervised Case: Unlabeled Development Data

Now suppose that our two competing models p_1 and p_2 have already been "trained," such that θ_1, θ_2 have been fitted by maximum-likelihood estimation to obtain θ_1^*, θ_2^*, but that we wish to leverage n additional *unlabeled data* examples $x_1, x_2, \ldots, x_n$ to accomplish the model selection task described in Section II-A above. Lacking the corresponding class labels $y_1, y_2, \ldots, y_n$ for these data, we thus seek to employ *automatically generated* labels $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ fitted respectively by maximum-likelihood under each of the two systems, such that we replace the conditional log-likelihood ratio of (1) by the generalized log-likelihood ratio

$$\sum_{i=1}^n \log \frac{p_1(x_i \mid \hat{y}_i; \theta_1^*)}{p_2(x_i \mid \hat{y}_i; \theta_2^*)} = \sum_{i=1}^n \log \frac{\max_y p_1(x_i \mid y; \theta_1^*)}{\max_y p_2(x_i \mid y; \theta_2^*)}. \tag{2}$$

Of course, maximum-likelihood labeling ("decoding") of Y given X incurs some error, and hence it is natural to ask under what conditions we can replace T_{lab} in the labeled-data model selection task of Section II-A with (2). Since this corresponds to the use of labels *taken as output from trained systems*—i.e., estimated under each of the two competing models p_1 and p_2—this procedure will inevitably suffer from misclassification errors with respect to the estimated labels; if systems p_1 and p_2 exhibit reasonable performance, however, the corresponding marginal error rate ϵ will be small. In the limit as ϵ tends to zero, of course, we recover precisely the setting of labeled data encountered in Section II-A above.

For the case of small but nonzero ϵ, and assuming now that the true data-generating model is either p_1 or p_2, we show below that a principled model selection procedure may obtained by adapting results from the labeled-data setting as follows. Each individual likelihood ratio $p_1(x_i \mid \hat{y}_i; \theta_1^*)/p_2(x_i \mid \hat{y}_i; \theta_2^*)$ will instead be *censored*, by bounding its range from above and below in order to limit the influence of misclassification

errors on the overall model selection procedure. In the limit, as we will see, this recovers the well-known nonparametric *sign test*, which simply tabulates for every $i = 1, 2, \ldots, n$ the sign of each log-likelihood ratio, rather than its actual value. As we formulate in Section II-C below, this approach sacrifices a degree of statistical efficiency for enhanced robustness, which in turn enables the influence of errors in the set $\{\hat{y}_i\}$ of automatically generated labels to be limited.

Not only is this approach intuitively reasonable, but it is also provably optimal in a minimax sense, as we now describe. To account for the misclassification errors induced by automatic labeling, we model the consequence of this inexact labeling procedure by replacing the exact conditional densities $p_1(x \,|\, y)$ and $p_2(x \,|\, y)$ with *mixtures* of these densities and "contaminating" distributions that represent the aggregate effects of misclassification. The misclassification error rate $\epsilon \ll 1$ moreover serves as the mixture weight for each respective contaminating density—the so-called ϵ-contaminated case [2].

Rather than seeking to determine these contaminating distributions directly, it is natural to ask if there exists a *least favorable* case: a form of contamination that, for fixed ϵ, would serve to maximize the probability of selecting the incorrect model p_1 or p_2. The answer is affirmative: Amongst all possible contaminating densities, we are guaranteed that a *least favorable* pair exists whenever the likelihood ratio $p_1(x \,|\, y)/p_2(x \,|\, y)$ is monotone and ϵ is small enough to ensure that the corresponding sets of admissible ϵ-contaminated mixtures remain disjoint.

In this case, a result obtained by Huber [2, Theorem 3.2] in the context of robust statistics may be applied to show that, to minimize this maximal risk of an error in model selection, it suffices to consider a specific form of contamination of p_1 by p_2, and vice-versa. The precise mixture form required by Huber's result is obtained by partitioning the range space of p_1 and p_2 in a manner that depends on $b > a > 0$ as follows:

$$\tilde{p}_1(x \,|\, \cdot) = (1-\epsilon) \cdot \begin{cases} p_1(x \,|\, \cdot) & \text{whenever } p_1 > a p_2, \\ a p_2(x \,|\, \cdot) & \text{otherwise;} \end{cases}$$

$$\tilde{p}_2(x \,|\, \cdot) = (1-\epsilon) \cdot \begin{cases} p_2(x \,|\, \cdot) & \text{whenever } p_2 > b^{-1} p_1, \\ b^{-1} p_1(x \,|\, \cdot) & \text{otherwise.} \end{cases}$$

A likelihood ratio test based on $\tilde{p}_1/\tilde{p}_2$ is thus seen to yield

$$\frac{\tilde{p}_1(x \,|\, \cdot)}{\tilde{p}_2(x \,|\, \cdot)} = \begin{cases} a & \text{if } p_1/p_2 \leq a, \\ p_1(x \,|\, \cdot)/p_2(x \,|\, \cdot) & \text{if } a < p_1/p_2 < b, \\ b & \text{if } p_1/p_2 \geq b, \end{cases}$$

and hence we have arrived at the minimax test for the case of ϵ-contaminated densities p_1 and p_2—a test based on likelihood ratio evaluations censored from below at a and above at b.

As noted by Huber, the limiting case occurs when ϵ is sufficiently large that the sets of ϵ-contaminated mixture densities $\tilde{p}_1, \tilde{p}_2$ cease to be disjoint, and begin to overlap; in our setting, this corresponds to the limit as a and b both approach unity. As a and b both approach unity, the log-likelihood ratio reflects only which term of the comparison is larger, yielding the *sign test* for model selection as described above:

$$T_{\text{unlab}} := \# \left\{ i : \log \frac{p_1(x_i \,|\, \hat{y}_i; \theta_1^*)}{p_2(x_i \,|\, \hat{y}_i; \theta_2^*)} > 0 \right\}. \tag{3}$$

This test statistic is distributed as a sum of n Bernoulli trials whenever the unlabeled examples $x_1, x_2, \ldots, x_n$ are independent and identically distributed, and is hence a binomial random variable. As such, we obtain a concrete directional test for model selection in the semi-supervised setting, in a manner that generalizes the supervised setting of Section II-A above.

As in the supervised case, we may fix a significance level α and determine a corresponding critical value k_α according to the binomial distribution with parameters n and p, where $p = 1/2$ under the null hypothesis of model equivalence. For a one-sided upper-tail test of size α, we reject $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ if $T_{\text{unlab}} > k_\alpha$, where k_α is the smallest integer such that $\sum_{k=k_\alpha}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n \leq \alpha$; reversing this inequality and summing from zero to k_α yields the corresponding one-sided lower-tail test. For a fixed alternate with $p \neq 1/2$, the corresponding probability of correct selection is given by $\sum_{k=k_\alpha+1}^{n} \binom{n}{k} p^k (1-p)^{n-k}$. The sign test has many appealing properties; we next investigate its statistical efficiency in this context, and refer the reader to [9] for other results.

C. Analysis: Comparing Statistical Efficacy and Efficiency

To summarize the results of [2] and [6] as they apply to our discussion of model selection above, the best test in the case of *labeled* development data accumulates the log-likelihood ratios of each example x_i given its correct label y_i, while in the case of *unlabeled* development data the corresponding minimax test accumulates the *signs* of these ratios when evaluated with respect to each automatically generated label $\hat{y}_i$. To compare the statistical efficacy of these two testing procedures, we may compute their *asymptotic relative efficiency* under general assumptions regarding the limiting distributions of (suitably standardized versions of) test statistics T_{lab} of (1) and T_{unlab} of (3) obtained under the null hypothesis.

Asymptotic relative efficiency expresses the limiting ratio of sample sizes necessary for two respective tests to achieve the same power and level against a common alternative; if one test has an asymptotic efficiency of 50% relative to another, then the former requires twice as many samples (in the large-sample limit) to achieve the same performance. Its computation requires knowledge of the asymptotic distributions of both test statistics under the null hypothesis, as we now describe.

Recall that when comparing strictly non-nested models using labeled data, a limit theorem holds under the null; let $f(\cdot)$ denote the associated density function, with corresponding variance σ^2. The so-called *efficacy* of the labeled-data test is in turn given by $1/\sigma$ under suitable regularity conditions, with that of the unlabeled-data sign test given by $2f(0)$ when T_{unlab} is appropriately standardized [9].

The corresponding asymptotic relative efficiency is in turn given by the squared ratio of test efficacies, which evaluates to the quantity $[2\sigma f(0)]^2$. This result implies that when T_{lab} is asymptotically Normal, the sign test corresponding to (3) is only $2/\pi \approxeq 64\%$ as efficient as the labeled-data test corresponding to (1), since $(2\sigma/\sqrt{2\pi\sigma^2})^2 = 2/\pi$. We may in fact generalize this result slightly by following the analysis of [10], and considering the so-called generalized Gaussian

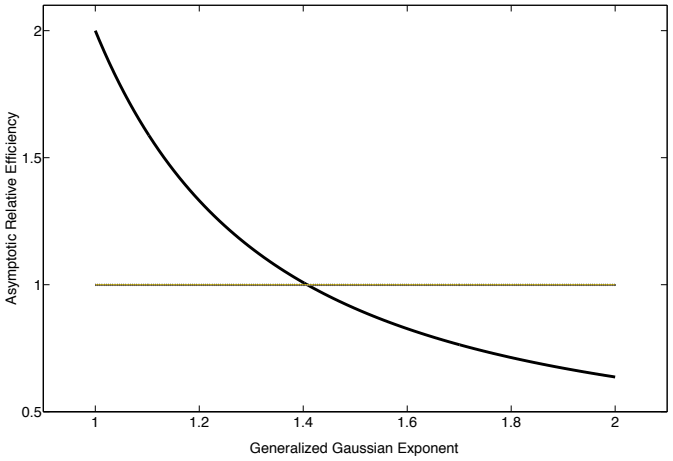

Fig. 1. Asymptotic relative efficiency of tests in the semi-supervised versus supervised settings, when the test statistic of the latter converges to a generalized Gaussian distribution with exponent p between 1 (Laplacian) and 2 (Normal). The horizontal line divides the range of p into cases for which the sign test is less efficient than the conventional likelihood ratio test, as in the case of the Normal, and vice-versa.

distribution with location parameter μ and scale parameter σ:

$$f_p(x) = \frac{1}{2\sigma\zeta(p)^{1/p}\,\Gamma(1+1/p)} \exp\left[-\tfrac{1}{\zeta(p)}\left(\frac{|x-\mu|}{\sigma}\right)^p\right].$$

Here $\Gamma(\cdot)$ is the Gamma function, $\zeta(p) = [\Gamma(1/p)/\Gamma(3/p)]^{p/2}$, and exponent $1 \leq p \leq 2$ allows us to interpolate between the Laplacian ($p = 1$) and Normal ($p = 2$) densities.

If we thus consider the expression $[2\sigma f_p(0)]^2$ for asymptotic relative efficiency, it follows from the relation $\Gamma(1 + 1/p) = \Gamma(1/p)/p$ that, as a function of exponent $p \in [1, 2]$, the asymptotic relative efficiency for the case of a generalized Gaussian distribution having exponent p is $p^2\Gamma(3/p)/[\Gamma(1/p)]^3$. This result is illustrated in Figure 1, which confirms that, were the asymptotic distribution of T_{lab} to approach a Laplacian density with $p = 1$, rather than a Normal with $p = 2$, the sign test would be twice as efficient in the large-sample limit.

D. Selecting from Amongst $k > 2$ Competing Models

As demonstrated above, the case of two competing hypotheses yields theoretical performance guarantees; however, in practice it is often necessary to select from amongst $k > 2$ models. While optimality is no longer necessarily retained [2], this problem is of sufficient practical interest to have generated a large contemporary literature in machine learning [11], [12].

Of the many approaches described in, e.g., [11], [12], several feature pairwise comparisons: in the so-called "one vs. all" method, each model is assigned a real-valued score relative to all others, and the model with the highest overall score is selected. Other possibly approaches include "tournament-style," following initial pairwise comparisons, or the case of all possible $\binom{k}{2}$ pairwise comparisons.

The latter approach has been suggested in [13] for the case of the sign test, and currently remains common practice within the machine learning community, despite multi-class procedures tailored to specific learning methods [11]. As such, we employ it to select amongst competing pronunciation models in our experiments below.

III. Application: Selecting Pronunciation Models

As a prototype application of the semi-supervised model selection approach derived in Section II, we now consider the task of evaluating candidate pronunciations of spoken words in large-scale speech processing tasks. To select amongst competing pronunciations, we consider two speech recognition systems that differ only in the pronunciation of a particular word, and show how to employ both the conventional test of (1) using *transcribed* audio data, and the sign test of (3) using *untranscribed* audio data.

A. Motivation for Semi-Supervised Pronunciation Selection

The selection of pronunciation models is crucial to several speech processing applications, including large-vocabulary continuous speech recognition, spoken term detection, and speech synthesis, each of which requires knowledge of the pronunciation(s) of each word of interest. In this setting, a set of admissible pronunciations forms what is termed a *pronunciation lexicon*, which comprises mappings from an orthographic form of a given word (e.g., *tornados*) to a phonetic form (e.g., /t er n ey d ow z/).

The conventional means of creating a pronunciation lexicon is to employ a trained linguist. However, as is the case with other examples requiring data to be hand-labeled by experts, this process is expensive, inconsistent, and even at times impossible, when individuals lack sufficiently broad expertise to create pronunciations for all words of interest [14]. In turn, several approaches for automatically *generating* pronunciations have been put forward [14], [15], [17], [19], [20], [21], and inevitably a model selection decision must be made to choose between candidate pronunciations. However, these approaches have themselves relied upon labeled training data, in the form of spoken examples of a given word and the corresponding orthographic transcripts.

In addition to the initial creation of a lexicon, pronunciation models are also necessary to maintain the vocabulary of speech processing systems over time: Although the pronunciation lexicon for a given system is created for as large a vocabulary as possible before deployment, this lexicon must be extended over time to incorporate *out-of-vocabulary* words. Such terms can be new words or names that come into common usage, rare or foreign words, or simply words not deemed significantly important at the time a system's lexicon was constructed. Dynamically adjusting to changing vocabularies thus requires the generation of *new* pronunciations over time, thereby reinforcing the need for an efficient and effective means of automatically selecting from amongst candidate pronunciations [22], [23], [24].

B. Methods for Selecting a Pronunciation Model

Much effort to date has been focused in the area of automatic pronunciation modeling—i.e., grapheme-to-phoneme or letter-to-sound rules. Previous work, including [14] and [15], has attempted to simultaneously generate a set of pronunciations and select between them. Also, work including [16] augments the possible pronunciations by building a larger

Word	Candidate Pron.	Reference Pron.
guerilla	g ax r ax l ax	g ax r ih l ax
guerilla	**g w eh r ih l ax**	
tornados	**t er n ey d ow z**	t er n ey d ow z
tornados	t ao r n ey d ow s	t ow r n ey d ow z

TABLE I
EXAMPLES OF CANDIDATE AND REFERENCE PRONUNCIATIONS

phone network to select the pronunciation. Additional resources are typically required, including existing pronunciation lexica [14], speech samples [19], [20], linguistic rules [21], or a combination of these. The focus of previous work has been on pronunciation variation [14], [19] or on common words [15], [17]. Note that in practice, other concerns may dictate choices between competing pronunciations, such as the scenario considered in [18], while highlighting the trade-offs between word accuracy and overall word error rate (WER). In the current setting, however, we are agnostic as to how the pronunciations are generated; our goal is simply to choose between them.

To this end, consider the setting in which we have example utterances $\{x_1, x_2, \ldots, x_n\}$, their corresponding transcripts $\{y_i\}$, and two "trained" speech recognition systems $p_1(\cdot)$ and $p_2(\cdot)$ that are identical (i.e., conditioned on the same parameters) *except* that for one word, models p_1 and p_2 use different pronunciations, say θ_1^* for $p_1(\cdot;\theta_1^*)$ and θ_2^* for $p_2(\cdot;\theta_2^*)$. This corresponds to the case of *strictly non-nested* models outlined in Section II. We subsequently describe and compare a supervised and semi-supervised method to select between candidate pronunciations θ_1^* and θ_2^*, and hence between models p_1 and p_2, in settings where candidate words are analyzed one at a time (as opposed to comparing entire pronunciation lexicons).

1) Supervised Selection of Pronunciations: The conventional mechanism for choosing between reference pronunciations of a word, examples of which are shown in Table I, is to acquire spoken utterances that contain the word, along with an orthographic transcription of the utterances, and compute a forced alignment of the acoustic waveform data to the transcripts, first using one pronunciation and then using the other [14], [15], [20], [21]. The pronunciation that is assigned a higher (Viterbi maximum likelihood) score during alignment is then chosen. For each word there are a fixed number of candidate pronunciations, with at least one (e.g., *guerilla*) reference pronunciation per word, although there may be several (e.g., *tornados*).

Cast in the notation of Section II, the conventional *supervised* method of pronunciation selection proceeds as follows:

1) Use the sequence of words comprising reference transcription $\boldsymbol{y}_i^{(\mathrm{ref})}$ for utterance $X_i = x_i$ to compute the log-likelihood ratio

$$\Lambda_i(x_i|\theta_1^*, \theta_2^*, \boldsymbol{y}_i^{(\mathrm{ref})}) = \sum_{y \in \boldsymbol{y}_i^{(\mathrm{ref})}} \log p_1(x_i|y;\theta_1^*) - \sum_{y \in \boldsymbol{y}_i^{(\mathrm{ref})}} \log p_2(x_i|y;\theta_2^*);$$

2) Use the n utterances to form T_{lab} and test as follows:

$$T_{\mathrm{lab}} = \sum_{i=1}^{n} \Lambda_i(x_i|\theta_1^*, \theta_2^*, \boldsymbol{y}_i^{(\mathrm{ref})}) \underset{\mathcal{H}_2}{\overset{\mathcal{H}_1}{\gtrless}} \tau_{\mathrm{lab}}; \quad (4)$$

3) Decide between $\mathcal{H}_1$ (model/pronunciation θ_1^*) and $\mathcal{H}_2$ (model/pronunciation θ_2^*) based on the difference in conditional likelihood evaluations, given forced-alignment reference transcripts, as indicated in (4).

2) Semi-Supervised Pronunciation Selection: The conventional method of pronunciation selection described above requires *transcribed audio data* whose production is a difficult, time-consuming, and laborious task. In many applications, external information can potentially alleviate the need for transcriptions by identifying recorded speech segments that are a priori likely to contain instances of a given word, which in turn may be used to select between candidate pronunciations. Examples include news items and television shows, each of which provides a rich source of *untranscribed* speech that could serve to improve the selection of pronunciations.

It is furthermore often the case that, while a transcript corresponding to spoken examples of a word is unavailable, we may have some knowledge that it has occurred in a particular audio archive. For example, we may know from weather records that a broadcast news episode recently aired about natural disasters, giving us a degree of confidence that instances of words like *tornados* are likely to appear. We may not know where or how many times such a word occurs in a particular audio segment, but we can still use the entire broadcast to help us choose between candidate pronunciations for *tornados*, examples of which are given in Table I.

In the *absence* of labeled examples we proposed to use the recognition system outputs themselves—unconstrained by any forced alignment or reference transcript—to select between candidate pronunciations. Each speech recognition system is run on every candidate data segment likely to contain a given word of interest, and from these results the corresponding acoustic likelihoods are evaluated with respect to the entire data set, leading to the selection of the candidate pronunciation yielding the highest overall likelihood.

Recalling our notation for the competing models $p_1(\cdot;\theta_1^*)$ and $p_2(\cdot;\theta_2^*)$, with corresponding pronunciations θ_1^* and θ_2^*, this semi-supervised approach proceeds in analogy to the labeled-data setting as follows:

1) Form the automatically generated word sequences $\hat{\boldsymbol{y}}_i^{(\theta_1^*)}$ and $\hat{\boldsymbol{y}}_i^{(\theta_2^*)}$ for each utterance $X_i = x_i$:

$$\hat{\boldsymbol{y}}_i^{(\theta_1^*)} = \operatorname*{argmax}_y p_1(y|x_i;\theta_1^*)$$
$$\hat{\boldsymbol{y}}_i^{(\theta_2^*)} = \operatorname*{argmax}_y p_2(y|x_i;\theta_2^*),$$

and use $\hat{\boldsymbol{y}}_i^{(\theta_1^*)}, \hat{\boldsymbol{y}}_i^{(\theta_2^*)}$ to compute the log-likelihood ratio

$$\Lambda_i(x_i|\theta_1^*, \theta_2^*) = \sum_{y \in \hat{\boldsymbol{y}}_i^{(\theta_1^*)}} \log p_1(x_i|y;\theta_1^*) - \sum_{y \in \hat{\boldsymbol{y}}_i^{(\theta_2^*)}} \log p_2(x_i|y;\theta_2^*);$$

2) Use the n utterances to form T_{unlab} and test as follows:

$$T_{\text{unlab}} = \#\{i : \Lambda_i(x_i|\theta_1^*, \theta_2^*) > 0\} \underset{\mathcal{H}_2}{\overset{\mathcal{H}_1}{\gtrless}} \tau_{\text{unlab}}; \quad (5)$$

3) Decide between $\mathcal{H}_1$ (model/pronunciation θ_1^*) and $\mathcal{H}_2$ (model/pronunciation θ_2^*) based on the number of log-likelihood ratios that evaluate to be positive, as indicated in (5).

IV. Large-Scale Experimental Validation

We now present an experimental validation of the semi-supervised model selection approach presented in the preceding sections, consisting of selecting between candidate pronunciations in the context of three prototypical large-scale speech processing tasks. For each of 500 different words, forced alignment and recognition outputs were produced for every pair of pronunciation candidates. Recognition was performed on an hour of speech for every word and each corresponding candidate, making sure to include somewhere in the data to be recognized the same speech utterances that were used in the forced-alignment setting, yielding a total of 1000 hours of recognized speech.

The quality of the selected pronunciations was then evaluated in three different ways: through decision-error trade-off curves for spoken term detection, phone error rates relative to a hand-crafted pronunciation lexicon, and word error rates for large-vocabulary continuous speech recognition. All experiments were conducted using well-known data sets, and state-of-the-art recognition, indexing, and retrieval systems.

A. Methods and Data

In order to evaluate the performance of semi-supervised pronunciation selection and its suitability for a variety of applications (e.g., recognition, retrieval, synthesis), and for a variety of word types (e.g., names, places, rare/foreign words), we selected speech from an English-language broadcast news corpus and identified 500 single words of interest. Common English words were removed from consideration, to ensure that words of interest would often be absent from lexicons, and thus would require pronunciation selection (e.g., *Natalie*, *Putin*, *Holloway*), and all words of interest featured in at least 5 acoustic instances. The selected words of interest were verified to be absent from the recognition system's vocabulary, and all speech utterances containing these words were removed from consideration during the acoustic model training stage.

For each word of interest, two candidate pronunciations were considered, each of which was generated by one of two different letter-to-sound systems [25]; furthermore, the 500 chosen words all had the property that the two letter-to-sound systems produced different pronunciations for them. For all subsequent experiments in semi-supervised pronunciation model selection, the sign test threshold τ_{unlab} was set at $\tau_{\text{unlab}} = n/2 + 1$, so that if more than half of the log-likelihood ratios evaluated to be positive, then the corresponding pronunciation model was chosen (i.e., a "winner-takes-all" approach). The threshold reflects our a priori belief of equally likely candidates, while enforcing our practical goal that one candidate or the other must be selected. The sensitivity to the threshold depends on the "distance" between models, as well as the number of observations. For the experiments in supervised pronunciation model selection, the threshold τ_{lab} was set at zero, so that the candidate with the higher log-likelihood was chosen.

Word	No. Samples	$\lvert T_{\text{lab}}\rvert$	$\lvert T_{\text{unlab}}\rvert$
Acela	8	151.92	4
afterwards	38	4846.52	31
Albright	247	34118.11	230
Barone	16	3011.04	12
Beatty	5	359.75	5
Iverson	21	1698.90	18
Peltier	12	741.12	9
Villanova	6	902.04	3

TABLE II
EXAMPLE WORDS AND THEIR ACCUMULATED TEST STATISTICS

To accomplish these experiments, a large-vocabulary continuous speech recognition (LVCSR) system was built using the IBM Speech Recognition Toolkit [26] with acoustic models trained on 300 hours of HUB4 data. Around 100 hours were used as the test set for recognition word error rate and spoken term detection experiments. The language model for the LVCSR system was trained on 400M words from various text sources. The LVCSR system's word error rate on a standard broadcast news test set RT04 (i.e., distinct from the 100 hours used for the test set employed below) was 19.4%. This LVCSR system was also used for lattice generation in the spoken term detection task. The OpenFST-based Spoken Term Detection system described in [27] was used to index the lattices and search for the 500 words of interest. For additional details regarding the experimental procedures and data sets, the reader is referred to [28].

B. Experimental Procedure

To summarize the experimental procedure, two alternative pronunciations are generated by two different letter-to-sound systems for each of a set of 500 selected words. We also have a *reference* pronunciation for these words from a hand-crafted pronunciation lexicon. We assume for the purposes of these experiments that the reference pronunciation is not available, and we set ourselves the task of choosing between two alternative pronunciations for each word, evaluated with respect to three different metrics, as will be discussed below.

The choice between the two pronunciations is made via either the supervised method of Section III-B1 (denoted *sup*) or the semi-supervised method of Section III-B2 (denoted *semi-sup*):

- *Sup* selects the candidate pronunciation based on supervised forced alignment with a reference transcript;
- *Semi-sup* selects the candidate pronunciation based on unconstrained (i.e., fully automatic) recognition.

Some example words of interest and their accumulated test statistics are shown in Table II. For each word, the number of true speech samples is listed, along with the accumulated log-likelihood ratios in accordance with (4), and the corresponding

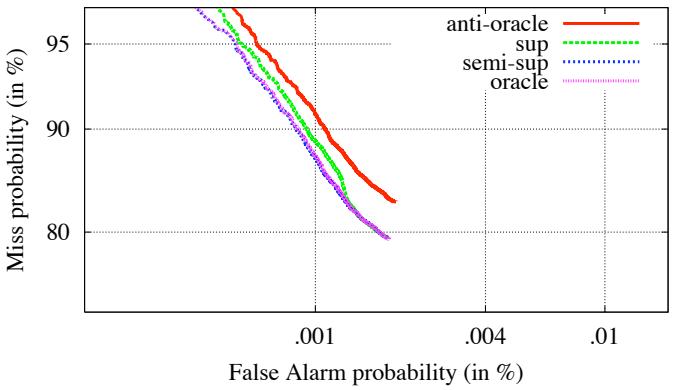

Fig. 2. Decision-error trade-off curves for a spoken term detection task [28], generated from 100 hours of speech data, using chosen pronunciations as queries to a phonetic/word-fragment index. Note that *semi-sup* and *oracle* overlap at nearly all operating points.

number of accumulated sign-test samples as per (5), in which the effect of likelihood censoring is apparent.

Additionally, we compare the methods described above with an *oracle* and an *anti-oracle*, defined with respect to the hand-crafted lexicon as follows:

- The *oracle* selects the candidate that has the *smallest* edit distance to a reference pronunciation of that word
- The *anti-oracle* selects the candidate that has the *largest* edit distance to a reference pronunciation of that word

To illustrate this notion, recall the earlier examples featured in Table I, which lists two words, each with two hypothesized pronunciations. In the case of these examples, the *oracle* pronunciation selection method would select the entries '/g ax r ax l ax/' and '/t er n ey d ow z/'.

C. Results

1) Spoken Term Detection: Experimental results from [28], showing the result of competing approaches to selecting between candidate pronunciations for purposes of spoken term detection, are shown in Fig. 2. Lattices generated by the LVCSR system for the 100-hour test set were indexed and used for spoken term detection experiments in the OpenFST-based architecture described in [27]; the chosen pronunciations were used as queries to the spoken term detection system. Results from the OpenFST-based indexing system were computed using standard formulas from the National Institute of Standards and Technology (NIST) and scoring functions/tools from the NIST 2006 spoken term detection evaluation. Note that the decision-error trade-off curves demonstrate that *semi-sup* performs better than the supervised method for detection at nearly all operating points.

2) Phone Error Rate (PER): This experiment measures which method—supervised or semi-supervised—selects pronunciations that have smaller edit distance to a reference pronunciation. Referring again to Table I as an example, if the bolded pronunciations had been selected based on the observed speech data, there would be 2 errors out of 6 phones with respect to the closest reference pronunciation for *guerilla*: delete /w/ and change /er/ to /ax/, resulting in a 33% PER; for *tornados*: 0% PER.

We note that while the supervised method requires a few *acoustic samples* of a word of interest, the semi-supervised method requires that a few instances of the word *be recognized*—correctly or incorrectly—by the LVCSR system. If insufficiently many instances are recognized, then a choice between alternative pronunciations cannot be made. Therefore, depending on the accuracy of the system, only a subset of the 500 words may be resolved (in the sense of having a pronunciation selected) by the semi-supervised method. Consequently, we employed three different levels of language model pruning to yield three levels of system quality, defined in terms of word error rate on the standard RT04 data set. The resultant error rates on the RT04 data set were 29.3%, 24.5%, and 19.4%.

We report the corresponding phone error rates in Table III, from which we observe that additional words are indeed resolved as system accuracy increases. By way of comparison, at the 19.4% WER system setting, the *oracle* method had a PER of 11.51%, and the *anti-oracle* had a PER of 27.2%. It may also be observed from Table III that, for those words which are resolved, the semi-supervised method (*semi-sup*) chooses candidates with smaller edit distance to reference pronunciations from a hand-crafted lexicon.

3) Large-Vocabulary Continuous Speech Recognition: As a final experiment, all four methods described in Section IV-B for selecting between candidate pronunciations were used to recognize 100 hours of speech that contained all 500 words of interest. Table IV shows a comparison of the results in terms of standard word error rates. Note that between the two alternative pronunciations, the one with the smaller phoneme edit distance to a reference pronunciation may not necessarily be the one that results in a lower word error rate. Overall, however, a range of about one-half of a percent of WER is observed between the best and worst candidates considered; note from Table IV that the supervised selection of pronunciations based on a forced alignment yields a slightly lower error rate in this instance than phoneme edit distance.

Finally, note that the semi-supervised method does as well as the supervised method. As shown in Table III, of the 449 words that were resolved, both the supervised method and the semi-supervised method selected the *same candidate* for 392 of them. Details of the remaining 57 words are presented in Table VI: Candidate pronunciations are listed in the second and third columns, with the better-performing candidate in bold, and columns 4 and 5 detail the *differing* errors due to selecting the candidate pronunciation *not in bold* in terms of substitution errors, and insertion/deletion errors. Many of the words where the methods chose different pronunciations do not impact word error rate—and hence neither is in bold—as the two candidate pronunciations are similar enough that neither results in a lower WER.

D. Selecting from Amongst $k > 2$ Competing Pronunciations

In practice it may be well necessary to compare more than two pronunciations for a given word. For example,

Method	System Quality (RT04 WER%)	No. Words Resolved	PER%	System Quality (RT04 WER%)	No. Words Resolved	PER%	System Quality (RT04 WER%)	No. Words Resolved	PER%
sup	29.3	359	13.00	24.5	390	13.66	19.4	449	14.50
semi-sup	29.3	359	12.64	24.5	390	13.19	19.4	449	13.87

TABLE III
PHONE ERROR RATES (PER) WITH RESPECT TO A HAND-CRAFTED LEXICON

Method	ASR WER%	No. Errors
anti-oracle	17.8	193,145
sup	17.3	187,772
semi-sup	17.3	187,424
oracle	17.4	188,517

TABLE IV
AUTOMATIC SPEECH RECOGNITION (ASR) WORD ERROR RATES (WER)

Method	ASR WER%	No. Errors
mw-anti-oracle	17.8	193,145
mw-sup	17.0	184,345
mw-semi-sup	17.0	184,297
mw-oracle	17.0	184,373

TABLE V
MULTI-WAY (MW) PRONUNCIATION SELECTION (3 PRONUNCIATIONS)

morphologically rich languages may dictate the consideration of $k > 2$ alternative pronunciations for a given orthographic form. To demonstrate that our techniques remain appropriate in this setting, we adopt here a strategy in which $\binom{k}{2}$ pairwise comparisons are performed for the case $k = 3$. In this approach, every unordered pair of candidate pronunciations is evaluated using the criteria described above for the *anti-oracle*, *sup*, *semi-sup*, and *oracle* methods. After all pairwise comparisons have been completed, the candidate chosen the greatest number of times is selected; as noted in Section II-D, a variety of alternative approaches are also possible.

For the results that follow, for each of the 449 words of interest, an additional *third* candidate pronunciation was considered, taken (as the last entry for a given word) from the reference pronunciation lexicon. Word error rate results for this three-way comparison are shown in Table V. The *anti-oracle* method WER remains the same as in the two-way case (Table IV), as every additional candidate had 0% PER, and by definition such candidates were not included in the *anti-oracle* set. In a similar fashion, the *oracle* set contained entirely reference pronunciations.

Relative to the earlier two-way comparison reported in Table IV, the *sup* and *semi-sup* sets here contained 288 and 301 new pronunciations, respectively. The remaining results summarized in Table V validate the trends observed in the two-way comparison, namely that *semi-sup* and *sup* perform comparably to each other, as well as to the *oracle*. Also, as expected, combining a third pronunciation of high quality resulted in lower error rates for all methods it affected.

V. DISCUSSION

In showing how censored likelihood ratios may be applied in the context of large-scale speech processing, we have developed in this article a semi-supervised method for selecting pronunciations using *unlabeled data*, and demonstrated that it *performs comparably* to the conventional supervised method. Empirical evidence in support of this conclusion was exhibited across three distinct speech processing tasks that depend upon pronunciation model selection: decision-error trade-off curves for spoken term detection, phone error rates with respect to a hand-crafted reference lexicon, and word error rates in speech recognition. We have observed these results to be consistent across many words of interest, based on extensive experiments using state-of-the-art systems and well-known data sets.

Note that there are limitations to this method, however, in the context of pronunciation selection. First, if neither candidate is ever recognized, the "unconstrained" recognition step required in the semi-supervised setting can fail to choose a candidate pronunciation for a word. Also, the approach requires having seen textual examples of the word of interest or words like it. This seems a reasonable requirement, given that a word comes into fashion by being widely noticed. Finally, false alarms in the recognition process may degrade performance—for example, if a word of interest sounds like common word—but our experiments to vary system quality indicated that this problem did not arise for the chosen words of interest in our setting.

In summary, the conventional supervised method for system-level model selection optimizes empirical performance on a labeled development set. Instead, we focused in this article on leveraging *unlabeled data* to choose amongst trained systems through likelihood-ratio-based model selection. We showed how to generalize the conditional likelihood framework through the use of *automatically generated* labels as a proxy for labels generated by human experts. We then answered the question of how well the resultant censored likelihoods are likely to perform, from both a methodological and an applied perspective.

As a final note, a current research direction of much interest to the speech community attempts to utilize untranscribed utterances for *self-training* of acoustic model parameters [4], [5]. While our main interest here was in the general problem of non-nested model selection using unlabeled data, an appealing direction for future work is to take these ideas forward within the acoustic modeling context.

VI. ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of colleagues at IBM Research and the use of their Attila speech recognition system [26], as well as support and the assistance of colleagues from a sub-team of the 2008 Center for Language and Speech Processing Summer Workshop at Johns Hopkins University, who helped to set up the necessary systems and plan experiments: Abhinav Sethy, Bhuvana Ramabhadran, Erica Cooper,

Murat Saraclar, and James K. Baker (co-leader). Also, we would like to acknowledge colleagues in the workshop for providing some of the pronunciation candidates, namely Michael Riley, Martin Jansche, and Arnab Ghoshal.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19–41, 2000.
[2] P. J. Huber, *Robust Statistics*. New York: John Wiley & Sons, 1981.
[3] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2006.
[4] F. Wessell and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 23–31, 2005.
[5] J. Ma, S. Matsoukas, O. Kimball, and R. Schwartz, "Unsupervised training on large amounts of broadcast news data," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2006.
[6] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, pp. 307–333, 1989.
[7] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, pp. 1–25, 1982.
[8] J. T. Kent, "Robust properties of the likelihood ratio test," *Biometrika*, vol. 69, pp. 9–27, 1982.
[9] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*. Springer, 2005.
[10] M. Kanefsky and J. B. Thomas, "On polarity detection schemes with non-Gaussian inputs," *J. Franklin Inst.*, vol. 280, pp. 120–138, 1965.
[11] A. C. Lorena, A. C. P. L. F. de Carvalho, and J. M. P. Gama, "A review on the combination of binary classifiers in multiclass problems," *J. Artif. Intell. Rev.*, 2009, in press.
[12] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, 2001.
[13] A. L. Rhyne and R. G. D. Steel, "A multiple comparisons sign test: All pairs of treatments," *Biometrics*, vol. 23, pp. 539–549, 1967.
[14] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labeled phonetic corpora," *Speech Commun.*, vol. 29, pp. 209–224, 1999.
[15] J. M. Lucassen and R. L. Mercer, "An information theoretic approach to the automatic determination of phonemic baseforms," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1984.
[16] D. Yu, M. Hwang, P. Mau, A. Acero, and L. Deng, "Unsupervised learning from users' error correction in speech dictation," in *Proc. Intl. Conf. Spoken Lang. Process.*, 2004.
[17] T. Vitale, "An algorithm for high accuracy name pronunciation by parametric speech synthesizer," *Computat. Linguist*, vol. 17, pp. 257–276, 1991.
[18] O. Vinyals, L. Deng, A. Acero, and D. Yu, "Discriminative pronunciation learning using phonetic decoder and minimum-classification-error criterion," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2009.
[19] B. Ramabhadran, L. R. Bahl, P. V. deSouza, and M. Padmanabhan, "Acoustics-only based automatic phonetic baseform generation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 1998.
[20] F. Beaufays, A. Sankar, S. Williams, and M. Weintraub, "Learning name pronunciations in automatic speech recognition systems," in *Proc. 15th IEEE Intl. Conf. Tools Artific. Intell.*, 2003.
[21] J. Teppermann, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, and S. Narayanan, "Pronunciation verification of children's speech for automatic literacy assessment," in *Proc. Intl. Conf. Spoken Lang. Process.*, 2006.
[22] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. 30th Ann. Intl. ACM SIGIR Conf.*, 2007.
[23] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. M. White, S. Khudanpur, H. Hermansky, and J. Cernocky, "Combination of strongly and weakly constrained recognizers for reliable detection of OOVs," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008.
[24] C. M. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence estimation, OOV detection, and language ID using phone-to-word transduction and phone-level alignments," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008.
[25] A. Sethy, M. Ulinski, S. Khudanpur, M. Riley, M. Jansche, A. Ghoshal, M. Saraclar, E. Cooper, D. Can, B. Ramabhadran, and C. White, "Web derived pronunciations for spoken term detection," in *Proc. 32nd Ann. Intl. ACM SIGIR Conf.*, 2009.
[26] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2005.
[27] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2008.
[28] C. M. White, A. Sethy, B. Ramabhadran, P. J. Wolfe, E. Cooper, M. Saraclar, and J. K. Baker, "Unsupervised pronunciation validation," in *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2009.

Term	*semi-sup*	*sup*	Differing Substitution Errors (No.)	Ins/Del
Ahern	**ey hh er n**	ae er n	ahern → upturn (3), apparent (2), hurry (1)	6
Aleve	**ae l iy v**	ax l eh v	(0)	1
anybody's	eh n iy b aa d iy z	eh n iy b ah d iy z	(0)	0
Asean	**ax s iy ih n**	ey s iy ih n	asean → asham (1) and → asean (1)	2
Assuras	ax sh uh r ih s	ax sh uh r ax z	(0)	0
Avi	ax v iy	ey v iy	(0)	0
Beatty	b iy ae t iy	**b ey t iy**	fabiani → beatty (1)	1
Bhuj	**b uw jh**	b uw zh	bhuj → pooch, boost, boots, chip, merge (1)	5
Canucks	**k ae n ax k s**	k ae n ah k s	canucks → connects (1) knox → canucks (1)	2
Cortese	**k ao r t ey z iy**	k ao r t eh z	cortese → he (2), tasty, daisy, taste (1)	5
Cuellar	**k w eh l er**	k y uw l er	cuellar → korea, out (1)	2
Dundalk	d ah n d ao l k	d ah n d ao k	(0)	0
Dura	**d uw r ax**	d uh r ax	dura → dora (1)	0
Durango	**d uh r ae ng g ow**	d uh r ae ng ow	durango → tarango (1)	1
freemen's	f r iy m eh n z	f r iy m ih n z	(0)	0
Gejdenson	g ey hh d ax n s ax n	g ey hh d ih n s ax n	(0)	0
Gough	**g ao f**	g ao	gough → goff (2), damien (1) schwarzkopf → gough (1)*	1
Grosjean	**g r ow s jh ih n**	g r ow jh iy n	grosjean → are, gross (1), on (1)*	1
Hadera	**hh ax d eh r ax**	hh ae d eh r ax	hadera → era, out (1)	2
Heupel	**hh oy p ax l**	hh y uw p ax l	heupel → goals (1)	1
Ilan	**ih l ax n**	ay l ax n	ilan → airline (1)	0
ilo	**ay l ow**	ih l ow	ilo → iowa, eyal, low (1)	0
Iverson	**ay v er s ax n**	iy v er s ax n	iverson → iverson's (14), the (1)	18
Jonbenet	**jh aa n b ax n eh t**	jh aa n b ax n eh	jonbenet → they (1)	1
Jurenovich	**jh uw r eh n ax v ih ch**	y uw r eh n ax v ih ch	jurenovich → renovate, renovation (3), average (2) jurenovich → events, pitch (2), want (1) jurenovich → against, batch, each, edge, irrelevant (1) jurenovich → edge, next, now, sh, tournaments (1)	22
Kmart	**k ey m aa r t**	k m aa r t	kmart → mart (9), answer (2), mark, out (1) has → kmart (1)	13
Lampe	l ae m p iy	l ae m p	(0)	0
liasson	**l y ae s ax n**	ae s ax n	liasson → hanson (1)	1
Likud's	l ih k ah d z	l ay k uw d z	(0)	0
Litke	**l ih k iy**	l ih t k iy	litke → the (1)	1
Lukashenko	**l uw k ae sh eh ng k ow**	l uw k ax sh eh ng k ow	lukashenko → i (1)	1
Marceca	**m aa r s ey k ax**	m aa r s eh k ax	marceca → because, cut (1) siegel → marceca (1)*	1
Matteucci	**m ax t ey uw ch iy**	m ae t uw ch iy	matteucci → see, to (1), matures (1)*	1
Menendez	**m eh n eh n d eh z**	m eh n aa n d ey	menendez → as (3) as → menendez (3)*	1
Milos	m ay l ow z	m ih l ow z	(0)	0
Mustafa	**m ah s t ax f ax**	m uw s t aa f ax	mustafa → some, sun (1)	1
Nasrallah	**n ae s r aa l ax**	n aa r aa l ax	nasrallah → rolla, drama, on (1)	3
Nhtsa	**n ey t s ax**	n t s ax	nhtsa → a, nitze (1)	2
Nkosi	**n k ow s iy**	ng k ow z iy	nkosi → cozy (1)	1
Orelon	ao r l aa n	ao r ax l aa n	(0)	0
Ouattara's	**w ax t ae r ax z**	aw ax t ae r ax z	ouattara's → tara's (1)	1
Pawelski	**p ao eh l s k iy**	p ao l s k iy	pawelski → belsky, ski (1)	2
Peltier	**p eh l t iy er**	p eh l t iy ey	peltier → tear (2), here, pepsi, years (1)	5
pre	**p r ax**	p r	pre → per (1)	0
Prodi	p r ax d iy	p r aa d iy	(0)	0
Sadako	**s ax d aa k ow**	s ae d ax k ow	sadako → got (1)	1
Schiavo	**s k y ax v ow**	sh ax v ow	schiavo → gavel, ski, elbow, oddball, on, out, will (1)	1
Schiavone	**s k y ax v ow n**	sh ax v aa n	schiavone → bony, bounty (2), a, money, it (1) schiavone → the, voting, about, donate, ioni, owning (1)	16
Schlossberg	sh l ao s b er g	sh l aa s b er g	(0)	0
Skurdal	s k er d ax l	s k er d aa l	scurbel → skurdal (1) skurdal → off (1)*	0
Taliban's	**t ae l ih b ax n z**	t ae l ih b ih n z	metallica → taliban's (1)	1
Thabo	**th aa b ow**	th ax b ow	thabo → and, tabor (2) m., problem (1) thabo → hobbled, in, tomlin, trouble, tumbling (1)	11
tornados	**t er n ey d ow z**	t ao r n ey d ow s	(0)	0
Yasir	y ax s iy r	**y aa s iy r**	yasir → oster (1)	1
Yugoslavs	**y uw g ow s l aa v z**	y uw g ow s l aa v s	(0)	0
Zhirinovsky	**zh ih r ih n ao v s k iy**	iy r ih n ao v s k iy	zhirinovsky → ski, skin, speak (1)	3
Zorich	**z ax r ih ch**	z ow r ih k	zorich → storage, h., is (2)	6

TABLE VI

WORDS WHERE THE METHODS DIFFER IN SELECTION. DIFFERING ERRORS LISTED CAUSED BY THE NON-BOLD PRONUNCIATION MARKED WITH AN ∗.