A Generative Student Model for Scoring Word Reading Skills

Joseph Tepperman* *Member, IEEE,* Sungbok Lee, *Senior Member, IEEE,*
Shrikanth Narayanan, *Senior Member, IEEE,* and Abeer Alwan, *Fellow, IEEE*

***Abstract*—This study presents a novel student model intended to automate word-list-based reading assessments in a classroom setting, specifically for a student population that includes both native and nonnative speakers of English. As a Bayesian Network, the model is meant to conceive of student reading skills as a conscientious teacher would, incorporating cues based on expert knowledge of pronunciation variants and their cognitive or phonological sources, as well as prior knowledge of the student and the test itself. Alongside a hypothesized structure of conditional dependencies, we also propose an automatic method for refining the Bayes Net to eliminate unnecessary arcs. Reading assessment baselines that use strict pronunciation scoring alone (without other prior knowledge) achieve 0.7 correlation of their automatic scores with human assessments on the TBALL dataset. Our proposed structure significantly outperforms this baseline, and a simpler data-driven structure achieves 0.87 correlation through the use of novel features, surpassing the lower range of inter-annotator agreement. Scores estimated by this new model are also shown to exhibit the same biases along demographic lines as human listeners. Though used here for reading assessment, this model paradigm could be used in other pedagogical applications like foreign language instruction, or for inferring abstract cognitive states like categorical emotions.**

***Index Terms*—reading assessment, pronunciation evaluation, Bayesian networks, children's speech, student modeling.**

I. Introduction

HOW does a teacher judge reading skills from hearing a child read words out loud? Each student's pronunciation is, of course, relevant evidence of reading ability. For individual words read in isolation, a new reader's skills are best demonstrated through various pronunciation-related cues, including their correct application of English letter-to-sound (or LTS) rules [28] - rules that describe the complex mapping from orthography to phonemes in English. Certain types of letter-to-sound decoding mistakes clearly testify to incorrect reading, such as the common tendency in young readers of English to make vowels "say" their own names [1]. Hesitancy and disfluency in decoding sounds from text are also indicative of underlying reading difficulties, and are manifested through suprasegmental pronunciation cues when reading aloud. What about the case of a child whose native language (or L1) is not English, or who speaks with a foreign accent? How would the teacher know if a particular word's variant in pronunciation is due to the child's inability to apply English LTS rules, or is simply typical of the child's pronunciation trends in general, when reading or speaking? A conscientious teacher, in an effort to remain unbiased in their assessments, would know what variants in pronunciation to expect of their student's accented speech (hopefully distinct from those caused by true reading errors), and would apply different assessment criteria based on what they know about the child's background and their own past experience teaching similar children.

Ultimately, in assessing reading skills on the word level, a teacher makes an inference as to their student's hidden cognitive state - the state of identifying a string of characters as the intended word, or not. This inference is based on the available evidence spoken by the child as well as what they know about the child's demographics, the target words in the test, and about accented speech and children in general. This is not, strictly speaking, the same as assessing the child's pronunciation (i.e. comparing their pronunciation with some predefined reference), since in some sense every speaker implicitly determines their own "correct" reference pronunciation when reading or speaking. A child with a foreign accent can, of course, be capable of reading English correctly and fluently. It is then the teacher's task to decide if a read pronunciation is consistent with the child's own personal reference - the child's phonological trends when speaking overall - or is the result of mis-applied LTS decoding.

The main benefit of using automatic reading assessment in the classroom is that it can free up teachers' time and energy to do what they do best: teach. A system that standardizes the regular assessments teachers would otherwise be conducting by hand can not only save them time, but can eliminate any potential teacher biases, provide a fine-grained pronunciation analysis, track long-term trends over a large number of students, and offer diagnoses of different types of reading difficulties, allowing teachers to focus on child-specific instruction and additional interventions. One goal of this paper is to demonstrate part of such a system, a new development toward automatically mimicking teacher judgments in list-based word reading tasks, a very common test format for evaluating young readers.

In theory, a student model for isolated word reading skills should be simple to model and replicate automatically. From a standard pronunciation dictionary we can have some notion of acceptable variants of a target word when read by native speakers of English. We can also assume a closed set of variants resulting from common LTS mistakes, and another

Manuscript submitted October 15, 2008. Revised and resubmitted July 24, 2009.

J. Tepperman (e-mail: jtepperman@rosettastone.com) is currently with Rosetta Stone Labs, Boulder, CO 80302, USA. S. Lee and S. Narayanan (e-mail: sungbokl@usc.edu, shri@sipi.usc.edu) are with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA. A. Alwan (e-mail: alwan@ee.ucla.edu) is with the Department of Electrical Engineering, University of California Los Angeles, 90095, USA.

set based on foreign accented speech - these could be either determined empirically or derived from rules well-documented by experts in child pedagogy. With acoustic models, we can choose which of the available variants best matches an unknown pronunciation. If the child is a native speaker, then only from recognizing a native variant should we infer acceptable reading. If the child is nonnative, then both the native and foreign-accented variants should indicate acceptable reading skills. In either case, automatically detecting any of the variants commonly coming from LTS mistakes will suggest that the child does not know how to read the target word.

For automatically assessing reading skills, this is an unrealistic and ineffectual model for many reasons. First off, it presumes that there will be no overlap among pronunciation categories, so that the source of the pronunciation will be obvious, and for many words this is not the case. For example, with the standard ARPAbet English phoneme set, the word "mop" may be pronounced as /M OW P/ both by children of any background who make the mistake of decoding a "long" vowel for the 'o' as well as by students with a Mexican Spanish accent - for children who speak with this accent, the source of this variant remains obscure[1]. Even for the target words for which there is no overlap in pronunciation categories, there are further problems with this method. The pronunciation variants of a given target are often so close as to remain difficult to reliably distinguish with state-of-the-art acoustic models [25] (or with human ears, as listening tests show [4]). Aside from this, segment-level pronunciation is only one observable cue to underlying reading ability - the accept/reject algorithm described above does not account for other evidence such as speaking rate or suprasegmental manifestations of fluency or hestiation. Knowledge of the child's background will change how this evidence is interpreted, and teachers are not necessarily conscious of how all these variables interact to inform their inference of a student's cognitive state. Clearly there is a very complex set of implicit decisions at work here. This is all just by way of outlining some of the many difficulties in creating an automatic method to judge reading skills.

In light of the complexity of the task and its cognitive modeling goals, we propose using a Bayesian Network to model the generative interactions among these many disparate cues in a framework that would represent how we hypothesize teachers conceive of a child's reading skills. Bayesian inference on a hidden cognitive state variable would then reflect the degree to which all the evidence and background knowledge of the child combine to color a teacher's judgment. The Bayesian framework is attractive for its flexibility in creating a hypothesized causal structure among variables, and has been used in past studies for student modeling [6], [7], [21]. The novel aspects of this work lie in the careful distinction between pronunciation and reading skills, the application to nonnative-speaking readers, and the use of pronunciation variant categories and student background in a unified generative student model. Here we intend to answer the following relevant questions about the chosen model and the perceptual data used to train it:

- How subjective are human assessments of reading ability?
- Does a subjective cognitive model perform better than a categorical decision based strictly on pronunciation?
- Which cues are most useful in making an automatic assessment, and how does inclusion of child demographic information affect the model?
- Can the model's generative structure be optimized for the training data, and does this improve automatic score performance?
- What biases, if any, does this automatic scoring method present, and how do these compare with biases in listeners' assessments?

Section II will give some background on the data and the modeling framework. A perceptual study on a subset of this corpus is described in Section III. Section IV explains how the feature set is estimated and Section V suggests how those features could be unified in a network structure. Results of experiments on various network structures and non-Bayesian classifiers will be reported in Section VI, and Section VII will interpret these results in light of the perceptual evaluations and the above questions about the model. Section VIII concludes with some ideas for future improvements and other potential applications for a student model such as this.

II. Background

A. The TBALL Project and its Context

This work was done as part of the TBALL (Technology-Based Assessment of Language and Literacy) project [2], a UCLA-USC-UC Berkeley collaboration in response to a growing demand for diagnostic reading assessments in US schools [19]. The project's goal is to develop components for a system that would administer tests and collect data in a classroom environment, automatically provide assessment scores, organize reports of the results for teachers, and recommend further assessments and interventions. TBALL concentrates on children in Kindergarten through Grade 2, younger than those of most related studies. Due to the demographic makeup of Los Angeles, one emphasis of the project has become the development of such a system specifically with speakers of Mexican Spanish or Chicano English dialects in mind [9]; dealing with dialectal variability is a challenge for robust automated processing of speech.

The most well-known studies in the area of automatic reading assessment [5], [12], [16], [29] typically use automatic scoring as one component of a computer reading tutor that provides feedback to children in real-time for pedagogical purposes. The focus of these past projects has (with a few exceptions) been on sentence- and paragraph-level reading and comprehension, in which a machine tutor will follow a child word-for-word and indicate if any reading errors are detected as the passage progresses. Reading errors in all of these studies are defined strictly in terms of segment-level pronunciation mistakes - the causal link from reading skills to pronunciation evidence is made to be quite direct, probably because these studies do not account for nonnative or accented children's

[1]Of course, the ARPAbet vowel /OW/ (the IPA diphthong /oʊ/) does not exist in Spanish, but when working exclusively with read English data, it is the closest vowel to the Spanish monophthong /o/.

speech, nor for the subjectivity in judgment that must occur in those cases. Hence, improvements in these methods have come from creative use of sentence decoding grammars and pronunciation variants (usually in terms of read word fragments), as well as expert knowledge of reading mistakes and appropriate acoustic model training and adaptation. This assessment of reading skills strictly in terms of speech decoding on a closed set of predetermined "correct" pronunciations will serve as a baseline method in this paper.

TBALL differs from these other studies in several ways. Our goals are focused on assessment (rather than tutoring) and we omit feedback to the student in favor of fine-grained results reported to the teacher. The TBALL assessment battery does include passage-level reading, but also many simpler tests that are common in reading assessment and are based on lists of items designed to measure, for example, a child's ability to identify isolated words on sight, to blend syllables or phonemes together, or to recite the names and sounds of English letters. The work in this paper is designed for automatic assessment of isolated words, and here we work exclusively with data elicited from the K-1 High Frequency (K1HF) and Beginning Phonic Skills Test (BPST) word lists [13], [22] - these will be described in more detail in Section IV. However, our modeling framework could easily be extended to other list-based assessments. Due to the wide range of "correct" pronunciations when reading, TBALL also treats reading assessment as a more complex task than simply deciding between close variants, as explained in Section I. Our advances in this paper and elsewhere have come through the use of multiple pronunciation categories (beyond those expected of native speakers or reading errors), features beyond the segment level, classification algorithms beyond automatic decoding of speech, and information about the child's background [2], [26], [27].

The student model proposed in this paper is inspired mostly by the Knowledge Tracing (KT) model of student knowledge demonstration and acquisition during a tutoring session [8]. KT posits that every answer a student gives when taking a tutor-guided test is either a direct demonstration of their actual knowledge (or lack thereof), or else a lucky guess or a temporary slip. The tutor's intervention can affect the child's inner knowledge state by actually teaching them, or the tutor can simply scaffold the child's answer without really imparting any understanding to them. A student model based on KT was used in [6] to generate automatic sentence reading scores in cases when the child's observed answers may be corrupted by errors in ASR results. That study showed each child's automatic scores from this model correlated well with their performance on standardized tests that they took at the end of the school year.

B. Bayesian Networks

A Bayesian Network is a graphical model that defines the joint probability of a set of variables $X_1, X_2, \ldots, X_n$ as

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \qquad (1)$$

where $Pa(X_i)$ are X_i's "parents" in the network - the variables on which we expect X_i to be conditionally dependent, either because there is a causal relationship between the parents and the "child," or because knowledge of the parents' values would influence our expectation of the child's [10]. This pre-defined conditional dependence allows us to simplify the inference of any variable's value given the others:

$$\begin{aligned} & P(X_1|X_2, X_3, \ldots, X_n) \\ = \; & P(X_1, X_2, \ldots, X_n)/P(X_2, X_3, \ldots, X_n) \\ = \; & \prod_{i=1}^{n} P(X_i|Pa(X_i))/\prod_{i=2}^{n} P(X_i|Pa(X_i)) \end{aligned} \qquad (2)$$

where X_1 is excluded from possible parents in this example's denominator. In a discrete classification task such as item-level reading assessment, the value of X_1 can be estimated as

$$\underset{x_1}{\operatorname{argmax}} P(X_1 = x_1|X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \qquad (3)$$

or x_1 can be taken as X_1's discrete value if

$$P(X_1 = x_1|X_2 = x_2, X_3 = x_3, \ldots, X_n = x_n) \geq T \qquad (4)$$

for some appropriate threshold T.

Bayesian Networks are versatile as automatic classifiers in many ways. They allow for both continuous and discrete variables, and can be trained on instances that are occasionally missing one or more of the variables' values (as is often the case for real-world data). Conditional dependencies among variables can be specified a priori or optimized based on a method such as the Tree-Adjoining Naive Bayesian (TAN) algorithm [10]. Dynamic Bayesian Networks (i.e. ones that track sequences of variables over time) have been used extensively to incorporate articulatory, prosodic, and audio-visual features in Automatic Speech Recognition (ASR) [14], [15]. Studies such as [7], [21] have also used Dynamic Bayesian Networks to track student knowledge acquisition in intelligent tutoring systems without ASR capability, and have been extended to automatically scoring sentence-level reading [6], though without the focus on nonnative speakers.

The probability distribution of each variable in the network is also an open design parameter, and must depend on the conditional dependencies built into the model. In this study we follow the methods used in the Bayes Net Toolkit [17], in which we implemented our models. We modeled continuous nodes as Gaussian distributions. With discrete parents, they take the form of a table of Gaussians - one for each combination of parent values. With continuous parents, these we modeled as linear Gaussians, in which the mean is a linear combination of the parents' values and the variance remains constant. Discrete variables with discrete parents are modeled simply as a table of conditional probability values over all combinations of parent values, but discrete variables with continuous parents are not so straightforward. One option is to artificially discretize the parents, though this usually results in poor parameter estimates and is sometimes computationally unfeasible. Instead we chose to model these variables as continuous multinomial logistic (or "softmax") distributions [18], which are commonly used in Neural Networks and

behave like soft decision thresholds between discrete values. For a continuous parent node X and a discrete child node R, the softmax distribution is defined as:

$$P(R=i|X=x) = \frac{exp(w_i x + b_i)}{\sum_j exp(w_j x + b_j)} \quad (5)$$

Here w_i represents the i^{th} decision boundary's normal vector, specifying the steepness (or softness) of the curve; b_i is simply this normal vector's offset. This node's parameters require iterative training.

Further details on the network used in this paper are presented in Sections IV and V.

III. Perceptual Evaluations

Since this study is concerned with automatic generation of subjective judgments, we organized some formal listening tests to determine the level of agreement between annotators as an upper-bound on the automatic results, and to determine any sources of bias or disagreement in human perception of reading skills. Five listeners were asked to give binary accept/reject scores to word-list recordings of twelve children collected with the TBALL child interface (an average of 14 items per child). One of the listeners (#1) was a professional speech transcriber, another (#2) was an expert in linguistics, second-language acquisition, and literacy assessment, and the remaining three were PhD students of speech technology with many hours of experience assessing data from the TBALL corpus. The stimuli from each child's list were presented to the listeners in chronological order so that they would hear the test items in the same sequence a teacher in the classroom would have heard them. To maximally use the data perhaps as an automatic scoring system would, they were allowed to listen to each item as many times as they wanted, and they could go back to previous items from the current child before moving on to the next child. Along with the word-level recordings, the listeners were given the intended target word for each test item, but were not told any background information about the children, so as to minimize any a priori bias in scoring. Their judgment of each item was then based on the child's pronunciation of that item, their performance on previous items, the relative difficulty of the item, anything the listener could infer about the child, and their past experience scoring other children. The recordings chosen were balanced for the following potential sources of bias:

- gender
- L1 (English or Spanish)
- grade level (Kindergarten, 1st, or 2nd)

Inter-annotator agreement is reported in Table I in terms of item-level percent agreement and Kappa statistic, and on the word list level in terms of score correlation. If a pair of binary scores from two listeners X and Y for a word w is represented by $S^w_{x,y}$ then their estimated probability of agreement is defined as

$$P_A = \frac{count(S^w_{x=acc,y=acc}) + count(S^w_{x=rej,y=rej})}{W} \quad (6)$$

where W is the total number of words scored by both X and Y. The percent agreement is then $P_A \times 100$. The Kappa statistic is similar to the percent agreement, but accounts for the possibility of chance agreement between the two listeners. For example, if one of the listeners arbitrarily rated everying as "accept," then that might result in a high percent agreement but a low Kappa statistic. It is defined as

$$Kappa = \frac{P_A - P_E}{1 - P_E} \quad (7)$$

where P_E is the estimated probability of agreement by chance, assuming the two annotators are acting independently:

$$\begin{aligned} P_E &= \frac{count(S^w_{x=acc}) \cdot count(S^w_{y=acc})}{W^2} \\ &+ \frac{count(S^w_{x=rej}) \cdot count(S^w_{y=rej})}{W^2} \end{aligned} \quad (8)$$

A list-level score for word list l made up of W_L words is defined as the percentage of "accept" scores in that list:

$$SL^l_x = \frac{count(S^w_{x=acc})}{W_L} \quad (9)$$

The correlation between two sets of list-level scores SL_x and SL_y is then calculated as

$$C(SL_x, SL_y) = \frac{\sum_l (SL^l_x - \mu^L_x)(SL^l_y - \mu^L_y)}{\sqrt{\sum_l (SL^l_x - \mu^L_x)^2 \sum_l (SL^l_y - \mu^L_y)^2}} \quad (10)$$

where μ^L_x is the mean of all list-level scores for listener X.

TABLE I
Inter-listener agreement in reading scores, in terms of percent agreement and Kappa for binary item-level scores, and correlation between overall list-level scores.

Listener #		*Listener #* 1	2	3	4
2	*% agreement*	87.9			
	Kappa	0.725			
	correlation	0.852			
3	*% agreement*	94.2	92.9		
	Kappa	0.845	0.844		
	correlation	0.917	0.925		
4	*% agreement*	88.3	91.9	96.5	
	Kappa	0.779	0.821	0.910	
	correlation	0.889	0.930	0.987	
5	*% agreement*	91.2	96.0	94.7	94.7
	Kappa	0.773	0.913	0.869	0.867
	correlation	0.911	0.990	0.944	0.950

Overall, agreement between all pairs of listeners was consistently high. The percent agreement ranged from 87.9% to 96.5% and the correlation was anywhere from 0.852 to 0.990. This indicates that the listeners generally agreed not only on how many items were acceptable for each child, but on which those acceptable items were. Similarly, high Kappa statistics indicated that the high percent agreement was not simply due to chance. These findings do set a high standard for automatic scores, but are consistent with other previous perceptual studies on this corpus [26], [27]. They show that subjectivity is indeed present in this scoring task, but that overall agreement is higher than we might expect for other tasks such as, for example, strict pronunciation scoring [3].

To measure bias in these scores, we compared the percentage of acceptable-rated items between demographic categories using a one-tailed z-test for difference in independent proportions. Only one listener (#2) gave significantly higher scores to Kindergarteners over 1st or 2nd graders ($p \leq 0.03$). Besides that, the only significant difference in scores was neither in gender nor grade level but native language - all five listeners gave native English speakers higher scores than nonnative ones ($p \leq 0.04$). Since this was common across all listeners and they were not told what each child's background was, this difference in proportions was probably not bias but simply indicates that the nonnative speakers chosen really were worse readers. However, the number of speakers is not large enough to draw any conclusions correlating native language with reading ability in general (though other studies like those reviewed in [11] have confirmed this link with larger sample sizes).

What about subjectivity between demographic categories - did these listeners ever agree more often for one type of student than another? The answer is yes, in every case of demographic duality. They agreed more for male students than for female students, for native English speakers than for nonnative speakers, and for Kindergarteners than for 1st or 2nd graders (all with $p \leq 0.0001$). This could indicate that there is more speech variability among female, nonnative, and older children, perhaps due to variations in child physical development or second-language acquisition. The proportion of acceptable items was not different based on gender or grade level - only the item-level agreement among listeners changed. This means that they assigned roughly the same number of acceptable scores to both categories, but their disagreement about which items were acceptable was higher for one category than the other. Note, though, that all agreement levels were still quite high regardless (89.6% and up, depending on demographic category), even when the difference in agreement between categories was significant.

Listener #1 provided reference scores for the rest of the data used in this study (beyond this 12-child subset). This listener had 93.6% agreement with the majority vote scores from the other four listeners (a somewhat objective reference), and the remaining four agreed with the others' majority scores 96.1% of the time. This difference in agreement proportions was significant only with $p \leq 0.08$, and so we regarded Listener #1's scores to agree with the majority vote reference about as much as those from the rest of the listeners did - enough for them to serve as a reliable reference. In light of studies such as [3], [24] which claim a maximum of 70% agreement or 0.8 correlation between human annotators when judging pronunciation, any of these listeners' scores could have served as a good reference because of the high inter-listener agreement overall. All automatic results will be reported in comparison with Listener #1's scores.

IV. Feature Estimation

Three types of variables for student modeling are used in this study - one type we call *Hidden* variables, another type we call *Evidence*, and a third we call *Underlying* features. Hidden variables are the scores for a child's cognitive state when reading that we intend to estimate automatically. These variables are literally hidden from us because a child may or may not know how to read a target word, and this knowledge state may or may not manifest itself in their pronunciation - i.e. they might know how to read but could accidentally say it wrong, or they might not know how to read but could guess the correct pronunciation. To estimate this hidden ability we can only gather the relevant cues and make an inference based on them, from a teacher's point of view. The Evidence variables are what a child would demonstrate at the time of a test and a teacher would observe firsthand when conducting it. These are cues related to the child's pronunciation and speaking style, and are derived from robust speech acoustic models and prior notions of pronunciation categories. Other types of Evidence such as visual cues might be useful in scoring, but they will not be investigated in this study. The Underlying features are extra information that may influence judgments of reading ability, but are known before the test and before its Evidence is elicited. These are things like the child's background, or information specific to the test items, such as their relative difficulty. All these variables will be summarized in Table II.

A. Hidden Variables

Our automatic word-level scoring model is essentially that of a two-step decision process: we first estimate values for the Evidence demonstrated by the child in reading a target word t out loud (Section IV-B below will discuss these in detail). This first step is really just a model for teacher perception of this Evidence. In the second step, all these Evidence features $E_t = e_t^1, \ldots, e_t^5$ are combined with the Underlying features $U_t = u_t^1, \ldots, u_t^6$ in a Bayesian Network to synthesize an item-level binary score for that item's Hidden variable, q_t - this second step is the student model proper, conceived from a teacher's perspective. An additional consideration in estimating this binary score is another hidden variable, r_{t-1}: the child's overall reading ability on the given word list. This overall skill level is modeled as a continuous variable defined as the percentage of acceptably-read items in the list up to and including $t-1$, a "running score" for test performance. It is used here on the assumption that, for example, if a child can read 16 out of 18 items in the list acceptably, then the 19th item will probably also be acceptable, though this could potentially propagate errors in the automatic scores. Strictly-speaking, r_{t-1} is not exactly a hidden variable since a pseudo-observed value for it is estimated from the inferred value of the hidden q_t and the previous overally score, r_{t-2}.

An automatic score through Bayesian inference on an item's hidden binary variable q_t is then estimated by evaluating the conditional probability that it is acceptable reading:

$$\begin{aligned} Sq_t &= P(q_t = acc. | E_t, U_t, r_{t-1}) \\ &= P(q_t = acc., E_t, U_t, r_{t-1}) / P(E_t, U_t, r_{t-1}) \end{aligned} \quad (11)$$

This leads to a continuous score for each item, which can be left as-is or thresholded and made binary for comparison with human binary scores. The overall reading score is initialized to $r_0 = 0$ and in test mode the running score that estimates it

is automatically updated based on Sq_t as the test progresses. The joint probabilities in Eqn. 11 will be specified with assumptions of conditional independence explained below, in Section V.

B. Evidence

Section I introduced the idea of comparing an unknown pronunciation to one or more predefined pronunciation lexica, each capturing a type of expected variation in pronunciation due to a specific source, like reading errors or foreign accent. We hypothesize that these pronunciation lexica are useful to guess the source of a pronunciation variant when judging words read out loud, and consequently to infer the child's reading ability that generated that source of variation. In other words, we can estimate the degree to which the pronunciation comes from the set of common letter-to-sound (LTS) decoding errors or from the child's expected phonological patterns, as this will help to link the pronunciation evidence to a cognitive state of word reading ability (though this is not always so clear a link, as Section I explains). This idea is similar to theories of Lexical Access in perception of spoken words [23], in which an incoming sequence of phonetic segments is compared with similar sequences that can form words in the listener's vocabulary.

For a speech observation O we estimate the likelihood that it belongs to a pronunciation lexicon λ_p as follows:

$$\begin{aligned} P(O|\lambda_p) &= \sum_n P(O, \lambda_p^n|\lambda_p) \\ &= \sum_n P(O|\lambda_p^n)P(\lambda_p^n|\lambda_p) \end{aligned} \quad (12)$$

where λ_p^n is the model for pronunciation n in lexicon p, and we assume O is conditionally independent of λ_p when λ_p^n is given. If we approximate this sum with its maximum, and we assume all pronunciations in the lexicon are equally likely, then Eqn. 12 reduces to

$$P(O|\lambda_p) \approx \max_n P(O|\lambda_p^n) \quad (13)$$

In this study we use three pronunciation lexica defined in terms of phoneme-level substitutions, insertions, and deletions: variants common to native English speakers (*NA*), variants common in Mexican Spanish accents (*SP*), and variants arising from predictable reading errors (*RD*). For example, if the target test word is "can" then most native children will pronounce it either as /K AE N/, or sometimes /K EH N/ when speaking quickly. A child with a Spanish accent might say /K AA N/, and one who makes the common LTS mistake of having the 'a' say its name might pronounce it as /K EY N/. These variants are determined based on rules observed in heldout data [32] as well as input from experts in child literacy.

These lexicon-based likelihoods were then used to estimate "Goodness of Pronunciation" (GOP) scores for each pronunciation lexicon:

$$GOP(\lambda_p) \equiv log(P(\lambda_p|O)) \quad (14)$$

The GOP score [30] is an estimate of the posterior probability of the acoustic models decoded for the lexicon p. From the approximation in Eqn. 13 we get:

$$log(P(\lambda_p|O)) \approx log\left(\frac{\max_n P(O|\lambda_p^n)P(\lambda_p)}{\sum_p \max_n P(O|\lambda_p^n)P(\lambda_p)}\right) \quad (15)$$

If we assume equal priors for all values of $P(\lambda_p)$ and we approximate the sum in the denominator by its maximum, this score becomes

$$GOP(\lambda_p) = log\left(\frac{\max_n P(O|\lambda_p^n)}{\max_{n,p} P(O|\lambda_p^n)}\right) \quad (16)$$

The numerator is the most likely sequence of phonemes in the lexicon, and the denominator is the most likely phoneme sequence in all possible lexica. In practice, this denominator is estimated using a "phoneme loop" recognition grammar allowing any possible sequence of phonemes to be decoded. Since the phoneme boundaries in the numerator and denominator might not match and the scores should be normalized by the length of each phoneme, the GOP for each numerator phoneme ph that begins at frame b and ends at frame e will take the following form:

$$GOP(ph) = \frac{log(P(O|ph))/(e-b)}{\sum_i \left[log(P(O|d_i)) \cdot (e_i - b_i)\right]/(e-b)} \quad (17)$$

where d_i is the i^{th} phoneme recognized by the loop, beginning at frame $b_i \geq b$ and ending at frame $e_i \leq e$. All acoustic models were monophone Hidden Markov Models with three states and 16 Gaussian mixtures per state. They were trained on standard 39-dimensional MFCC features taken from 19 hours of children's speech collected in classrooms as part of the TBALL project, of which only a small amount was annotated on the phoneme level to bootstrap automatic transcription of the remainder. Viterbi decoding was used to estimate the likelihood of the observed speech given each model.

In addition to these posterior scores for each of the three pronunciation lexica, we also used two more features as Evidence. One was the child's rate of speaking (*ROS*), defined as the number of phonemes read per second. The second was the maximum a posteriori pronunciation recognition result, a discrete variable for the pronunciation lexicon with the highest GOP score:

$$\hat{p} = \underset{p}{\operatorname{argmax}}\, P(\lambda_p|O) \quad (18)$$

If each pronunciation lexicon's posterior represents a teacher's perception of an unknown pronunciation's distance to the pronunciations in that lexicon, then this pronunciation recognition result represents which pronunciation lexicon a teacher would choose if they had to pick only one.

C. Underlying Features

There are many things that we might expect to affect a teacher's inference as to a child's reading ability. Some of these are related to what they may already know about the child in question. In this study we use the child's native language, gender, and grade level (K, 1st, or 2nd) as three discrete Underlying variables to inform our automatic assessments.

The vast majority of the children in this dataset were native speakers of either English or Spanish, so we chose to represent

TABLE II
SUMMARY OF EVIDENCE, HIDDEN, AND UNDERLYING VARIABLES USED IN THIS STUDY.

	symbol	*variable*	*cardinality*
E_t	e_t^1	rate of speaking: *ROS*	cont.
	e_t^2	native lexicon GOP: *NA*	cont.
	e_t^3	Spanish lexicon GOP: *SP*	cont.
	e_t^4	reading mistake lexicon GOP: *RD*	cont.
	e_t^5	lexicon of maximum GOP: $\hat{p}$	3
	r_{t-1}	list-level running score	cont.
	q_t	item-level score	2
U_t	u_t^1	item index	cont.
	u_t^2	word length	cont.
	u_t^3	word list	4
	u_t^4	native language: *L1*	2
	u_t^5	gender	2
	u_t^6	grade level	3

native language as a binary variable. Their native language was determined through questionnaires filled out by the children's parents, and some of them chose not to respond - in those cases, we left this variable's value unspecified in the Bayesian Network. The same went for the few children whose native language was something other than Spanish or English. Our pronunciation lexica only accounted for variability related to a native English, L.A. Chicano English, or Mexican Spanish accent, and children with another native language were too few to justify changing the cardinality of this variable. Several of the children were described as bilingual - equally native to both Engish and Spanish. These we tagged as Spanish-speaking because they could potentially have the influence of the only foreign accent we examined. The reader should keep in mind that, whatever the child's native language, we did not have an objective measure of the presence of a foreign accent in their speech, and moreover, children who are native Spanish speakers do not necessarily demonstrate any discernible foreign accent.

Other Underlying features that may potentially sway a child's reading ability are factors related to the design of the test itself. Some words are more difficult than others, and this difficulty may manifest itself in those words' Evidence and Hidden variables. In these experiments we used recordings from four different word lists - three lists of K1HF words and one BPST list (see Section II-A) - and BPST is for most items the more difficult list. The K1HF lists are made of high-frequency words that beginning readers have hopefully already read many times: "I," "like," "can," "see," etc. The BPST words are designed to elicit distinctions between phonemes, and so contain many minimal pairs of less-common words like "map" and "mop," or "rip" and "lip." Because of a potential difference in word list difficulty, we included the word list as a discrete variable of cardinality 4. For similar reasons, the length of each word (in characters) was included as a continuous variable to account for increased difficulty proportional to the word length. Lastly, the item index divided by the total number of items in the list was used as another continuous variable. This made sense because the BPST list

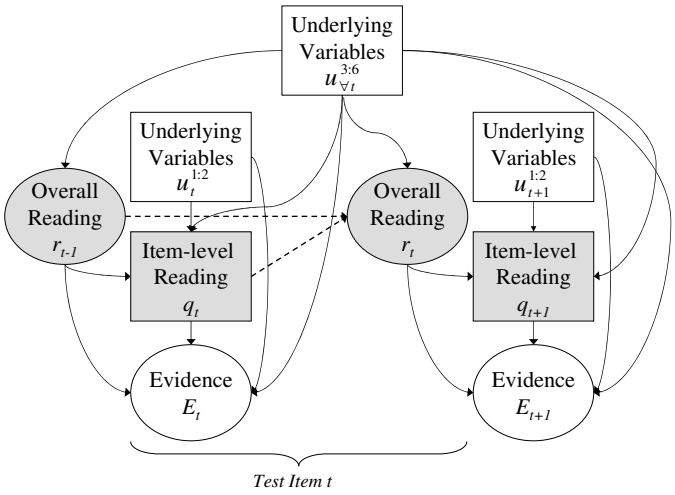

Fig. 1. Graphical illustration of the student model, over 2 test items. Shaded nodes denote hidden variables. The dashed lines are not probabilistic relations, but indicate how the overall score for item t is derived from the combined previous item and overall scores.

steadily increased in difficulty as the items progressed, and the item index also worked in conjuction with the overall score hidden variable r_{t-1}. For example, a running score after only 2 items should have less influence on the next item's binary score than the same running score after 15 items.

V. NETWORK STRUCTURE

Section IV described the feature set and how each variable's value was derived. Now we explain the conditional dependencies in our model among these variables, used in performing Bayesian inference on the item-level Hidden variable q_t, as in Eqn. 11: $P(q_t|E_t, U_t, r_{t-1})$.

A. Hypothesized Structure

To reiterate, the student model proposed in this paper takes what teachers perceive during the test (the Evidence, E_t), what teachers know beforehand about the student or the test (the Underlying variables, U_t), and the student's performance on prior test items as an estimate of their overall reading skill (the running score, r_{t-1}), and it uses all these things to construct a student model from a teacher's point-of-view that can be used to infer whether or not the student read the current test item acceptably, cognitively-speaking (i.e. not just in terms of pronunciation). Figure 1 shows graphically in a high-level Bayes Net structure what we hypothesize the conditional dependencies among these different types of variables should be, based on how we expect teachers conceive of reading ability and its demonstration. On these assumptions of conditional independence, Eqn. 11 simplifies to:

$$\begin{aligned} & P(q_t|E_t, U_t, r_{t-1}) \\ = \; & P(q_t, E_t, U_t, r_{t-1})/P(E_t, U_t, r_{t-1}) \\ = \; & \frac{P(q_t|U_t, r_{t-1})P(E_t|q_t, U_t, r_{t-1})P(r_{t-1}|u_{\forall t}^{3:6})P(U_t)}{P(E_t|U_t, r_{t-1})P(r_{t-1}|u_{\forall t}^{3:6})P(U_t)} \\ = \; & \frac{P(q_t|U_t, r_{t-1})P(E_t|q_t, U_t, r_{t-1})}{P(E_t|U_t, r_{t-1})} \end{aligned} \tag{19}$$

When performing Bayesian inference, these marginal probabilities are computed over an entire network using the junction tree algorithm of exact inference [17].

The item-level pronunciation Evidence, E_t, is a consequence of the student's hidden reading ability states (both item-level and overall running score), though how those states are manifested in all the Evidence is not deterministic. The hidden binary item-level state q_t is conditionally dependent on the overall skill level r_{t-1} estimated from the previous test items, as well as all the Underlying variables, U_t, since we would expect item-level inference to change based on the value of these parent variables. Note that this does not mean, for example, that the student's L1 is modeled as the *cause* of their reading ability - this is just one variable on which we would expect a teacher's inference into the student's cognitive state to depend. The overall reading skill, too, is modeled as a child of the Underlying variables but only those that apply globally ($u_{\forall t}^{3:6}$: word list, L1, gender, and grade level) rather than those that apply only to an individual test item ($u_t^{1:2}$: item index and word length). We also model the Evidence as a child of all the Underlying variables, so that the observed pronunciation features are conditionally dependent on both the student's hidden reading ability and other external factors such as word difficulty. The dotted lines from q_t and r_{t-1} to r_t do not denote probabilistic relations but rather show how the running score is updated at each item with the newly-inferred value of q_t. Table III shows the hypothesized child-parent relationships among all our variables in more detail.

Though we do not use any intelligent tutor feedback, our student model is in many ways analogous to that of Knowledge Tracing (KT) as introduced in Section II-A. The Hidden student cognitive state for each word item represents whether they know how to read the target word acceptably or not, and is equivalent to KT's "Student Knowledge" state for a particular skill. Both student models use observed "Student Performance" during the test (in our case, pronunciation Evidence) as a variable generated by the Hidden knowledge state. The Tutor Intervention variable in KT is similar to our Underlying features, which are in both models considered to be parents of the Hidden knowledge state and the observed Evidence. The novel aspect of our student model compared to KT and other similar models in [7], [21] is mainly in the feature set - our use of Underlying demographic and test item information as well as Evidence scores over several different pronunciation lexica is unique to this work.

TABLE III
HYPOTHESIZED PARENT-CHILD ARCS IN THE BAYESIAN NETWORK STUDENT MODEL. ONLY FOR PAIRS MARKED WITH AN 'X' IS THE COLUMN VARIABLE CONSIDERED CONDITIONALLY DEPENDENT ON THE ROW VARIABLE - ALL OTHERS ARE ASSUMED TO BE INDEPENDENT.

			Children						
					E_t				
					e_t^1	e_t^2	e_t^3	e_t^4	e_t^5
Parents			r_{t-1}	q_t	*ROS*	*NA*	*SP*	*RD*	$\hat{p}$
U_t	u_t^1	*item index*		X	X	X	X	X	X
	u_t^2	*word length*		X	X	X	X	X	X
	u_t^3	*word list*	X	X	X	X	X	X	X
	u_t^4	*L1*	X	X	X	X	X	X	X
	u_t^5	*gender*	X	X	X	X	X	X	X
	u_t^6	*grade level*	X	X	X	X	X	X	X
		r_{t-1}		X	X	X	X	X	X
		q_t			X	X	X	X	X

B. Structure Training and Refinement

The many hypothesized arcs in our graphical model (Fig. 1 and Table III) may result in a sub-optimal network for several reasons. First of all, some of the variables thought to be dependent may not in fact be, and modeling such "dependencies" would be useless. This would be a failure to follow the Occam's Razor principle of model succinctness, and would unnecessarily increase the computational complexity involved in estimating an automatic reading score. Beyond that, with finite training instances and an overly complex model there is always the possibility that true dependencies might not be estimated properly due to a dearth of training instances representative of all combinations of dependent variables.

For these reasons we propose an alternative forward-selection greedy search algorithm to refine the network structure of the hypothesized arcs. The algorithm begins with just one arc in the network representing a baseline dependency: the arc from q_t to the pronunciation lexicon recognition Evidence variable, $\hat{p}$. Then it proceeds in a random order to add each hypothesized arc individually, keeping an arc if it improves the likelihood of the training variables given the Bayesian Network. This process is looped until it has been shown that adding any remaining hypothesized arc will decrease the model likelihood. This method is also useful in that analysis of the refined network may reveal the necessary inter-variable dependencies of the data.

Because of occasionally missing data in some of the demographics variables, and the potential for some variables to be modeled as "softmax" distributions (for discrete features with continuous parents), all model parameters were estimated using the Expectation-Maximization (EM) algorithm that these conditions require. For a given Bayes Net with known structure S and initially random parameters θ^i, if we have observed features F and some missing data M, then each iteration of EM first estimates the expected value of the log-likelihood of the features given the model, with respect to the distribution of missing data given everything else:

$$Q(\theta, \theta^i) = E[logP(F, M|S, \theta^i)|F, S, \theta^i] \tag{20}$$

and then finds the new maximum likelihood estimate for the parameters:

$$\theta^{i+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^i) \tag{21}$$

The likelihood of the training set given the model - $P(F, M|S, \theta^i)$ - was defined as the log-likelihood after EM convergence, and convergence was defined as either 10 iterations of EM or the number of iterations required to make the following inequality true:

$$\frac{|LL(i) - LL(i-1)|}{\text{mean}\{|LL(i)|, |LL(i-1)|\}} < 0.001 \tag{22}$$

TABLE IV
DEMOGRAPHIC MAKEUP OF THE TEST DATA USED IN ALL EXPERIMENTS. TOTAL SPEAKERS MAY NOT ADD UP TO 189 SINCE SOME OF THIS INFORMATION WAS MISSING FOR CERTAIN STUDENTS FOR REASONS EXPLAINED IN SECTION IV-C.

	speakers	*word items*	*word lists*
Male	80	3950	278
Female	100	5051	349
English L1	71	3041	209
Spanish L1	82	4660	323
Kindergarten	35	882	104
1st grade	72	3854	253
2nd grade	82	4884	301

TABLE V
BASELINE PERFORMANCE USING ONLY THE GOP SCORE FOR THE *NA* LEXICON.

	Bayesian Network	*Support Vector Machine*
% agreement	83.60	83.60
Kappa	0.408	0.410
correlation	0.700	0.699

Here $LL(i)$ is the log-likelihood after iteration i. In most cases, EM on our data met this inequality within 3 or 4 iterations.

VI. EXPERIMENTS AND RESULTS

To paraphrase the overall questions first posed in Section I, our experiments with this new student model were intended to answer the following:

- Does the Bayes Net student model offer improvements over a baseline pronunciation-based paradigm for reading scoring?
- How useful are the novel features proposed here in making a reading score?
- Can we automatically learn the structure of the student model?
- How do the automatic scores compare with human scores, both in terms of agreement and bias?

The evaluation dataset used in all these experiments consisted of 6.85 hours of read words from 189 children - a total of 9617 word items from 658 word lists (an average of 15 items per list). These children were all distinct from those in the 19 hours of speech used for acoustic model training described in Section IV-B. The demographic makeup of this dataset is given in Table IV. Automatic scoring on the evaluation set was done using a five-fold crossvalidation procedure in which, for each fold, four-fifths of the speakers in the eval set were used to train and refine the student model which was then tested on the remaining one-fifth of the speakers. All variables were estimated using the acoustic models and background information as explained in Section IV.

A. Model Comparison and Baselines

Based on the methods in [5], [6], [12], [16], [29] of doing automatic reading assessment as a pronunciation recognition/verification task, we propose the following baseline. Here we assume that only one of the *NA* lexicon of pronunciations (those common to native-speaking readers) can qualify as a demonstration of acceptable reading ability, as past studies have done. The GOP score of the *NA* pronunciations, $P(\lambda_{NA}|O)$, then serves as the sole feature for our baseline experiments. Note that this baseline does not use scores from multiple pronunciation lexica: it is blind to child demographics as well as the prior knowledge of partitioning pronunciations into relevant categories.

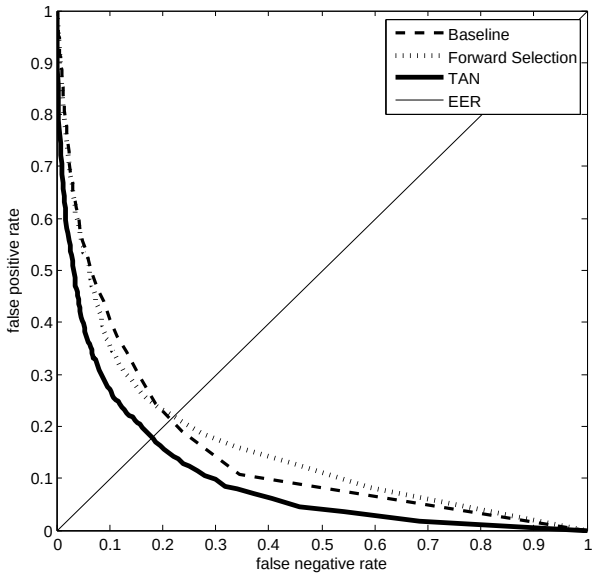

Fig. 2. DET curves over a varying threshold, for item-level binary classification on several different network structures. TAN and Forward Selection used the set of all features, in contrast to the Baseline.

Table V gives the baseline classification results (with the GOP score for lexicon *NA* as the only feature) for two contrastive classifiers: a simple Bayesian Network (Naive by default), and a Support Vector Machine (SVM) with Polynomial kernel function. A trained SVM specifies a hyperplane to separate a set of classes in a high-dimensional feature space [31]. It does this by optimizing the error bounds between the classes. A nonlinear discriminant function can be trained by using a nonlinear kernel function to transform the hyperplane. The SVM experiments are included here to compare the proposed generative student model to a purely data-driven discriminative classifier, with no structure reflecting a student's cognitive processes. The point of using single-feature baselines such as these is to show improvement in automatic reading scores with the addition of the novel aspects of this paper: the pronunciation Evidence based on multiple pronunciation lexica, the Underlying features, and the student model's network structure that unites them.

B. Network Structure Comparison

As explained in detail in Section V-B, the structure of the student model's network was automatically refined using a forward-selection procedure on the hypothesized conditional dependencies outlined in Section V. This refinement was done five times - once for each fold in the crossvalidation. Table VII gives the total number of times each hypothesized arc was

TABLE VI
ITEM-LEVEL % AGREEMENT AND KAPPA, AND LIST-LEVEL OVERALL SCORE CORRELATION BETWEEN AUTOMATIC RESULTS AND HUMAN READING SCORES.

Model			*all features*	*no* E_t	*no* U_t	*no* r_{t-1}
Bayesian Network	Forward Selection	*% agreement*	84.56	78.15	83.80	84.52
		Kappa	0.507	0.064	0.452	0.504
		correlation	0.750	0.436	0.702	0.750
	TAN	*% agreement*	87.42	81.22	86.51	86.01
		Kappa	0.616	0.322	0.590	0.576
		correlation	0.869	0.739	0.825	0.809
Support Vector Machine		*% agreement*	86.71	78.44	84.30	86.20
		Kappa	0.558	0.000	0.482	0.544
		correlation	0.824	-	0.698	0.794

selected for the final network, out of five crossvalidation folds. For comparison with this greedy search procedure, we also trained a network specified by the Tree-Adjoining Naive Bayes (TAN) structure learning algorithm [10]. A Naive Bayesian classifier assumes that all features are dependent only on the root classification node (in this case, the item level score q_t) and are independent of one another. The TAN algorithm results in a network structure that is similar to that of Naive Bayes, except it allows for each feature's probability distribution to be conditionally dependent on one other feature (selected according the Mutual Information criterion and not based on the cognitive model proposed in Section V). Because it is not available as part of the Bayes Net Toolkit software package, this algorithm was implemented using the Weka toolkit [31].

In these experiments, all item-level scores derived from a Bayesian Network model (of any structure) are defined as the argument that maximizes q_t in its conditional probability given all the features and the structure of the network: $P(q_t|E_t, U_t, r_{t-1})$. Since q_t is binary, this is equivalent to choosing $q_t = accept$ if

$$P(q_t = accept|E_t, U_t, r_{t-1}) \geq T \tag{23}$$

for $T = 0.5$, and choosing $q_t = reject$ otherwise. Item-level percent agreement and Kappa are reported in Tables V and VI, and these are shown alongside the correlation in list-level scores derived from the item-level ones according to Eqn. 9. Additionally, so as to get a sense for each structure's performance along a range of possible error rates, Figure 2 shows detection-error tradeoff (DET) curves for the item-level results over varying values of the threshold T in Eqn. 23. For comparison, the Baseline GOP score is also included in this Figure, with a varying threshold on the probability $P(\lambda_{NA}|O)$.

C. Feature Comparison

Are pronunciation features beyond simple native-accent pronunciation scores necessary when scoring reading automatically? This is part of what these experiments were intended to answer. If the answer is yes, then which of the new features (E_t, U_t, and r_{t-1}) are most useful in estimating q_t? To measure this, we redid the scoring experiments on the Bayes Nets and SVM, leaving each of these three feature subsets out. Table VI gives their item agreement and list-level correlation performance, to be compared directly with results obtained from the set of all features, and with the baseline results in Table V.

VII. DISCUSSION

A. Automatic Performance Comparison

The baseline results reported in Table V are almost identical for both the Bayes Net and SVM classifiers. This is as expected, considering they each used the same single baseline feature. On the item level, both had comparable agreement levels and apparently each one agreed with the human labels on the same items: the item-level matched-pair results were not significantly different using McNemar's test. Neither the Kappa agreement nor correlation of either baseline came close to the inter-listener agreement reported in Table I, though the 0.41 Kappa statistic meant that the item-level percent agreement results were well above chance levels.

According to Table VI, the best-performing classifier was the Bayesian Network with the TAN structure and all features included. Its item-level scores were significantly different from that of the second-place SVM (with McNemar's test and $p \leq 0.01$), and its correlation in list-level scores was also significantly higher than the SVM's (0.869 vs. 0.824 - a significant difference with $p \leq 0.01$). With the full set of features, all three classifiers significantly outperformed the baseline in terms of correlation and agreement, with $p \leq 0.05$ or better. Moreover, the TAN network's list-level correlation with Listener #1's scores surpassed the correlation between Listener #1 and Listener #2.

However, Figure 2 illustrates, over a range of thresholds, the similarity of the Baseline and the Forward Selection Bayes Net. Though at various points one curve may be closer to the origin than the other, they appear to intersect right at the equal error rate line. In contrast, the TAN curve is closest to the origin across all operating points. Though the Forward Selection method in Table VI did beat the single-threshold baseline as reported in Table V, the dominance of the TAN network in the results suggests that the the main source of improvements over the baseline proved to be the set of novel features proposed in this work and not the hypothesized network structure.

B. Automatic Structure Refinement

Though the Forward Selection structure learning algorithm did not perform as well as the simpler TAN method, the results of this selection procedure still merit some interpretation. According to Table VII, an average of 6 out of the 51 hypothesized arcs were excluded from the network in each fold of crossvalidation. This indicates that the hypothesized structure is for the most part an accurate model of the data's conditional dependencies. Interestingly, the dependencies excluded were generally ones that would have required variables to be modeled with softmax distributions - i.e. ones that linked continuous parents to discrete children. For example, the two continuous Underlying variables (word length and item index) were rarely ever allowed to be parents of q_t (the hidden binary reading variable) or $\hat{p}$ (the discrete recognition result). Similarly, the continuous-valued overall score r_{t-1} was generally not selected as the parent of either of those variables. Perhaps this illustrated a limitation in estimating the softmax distribution's parameters (which required EM), or in using the softmax distribution itself. It is also important to keep in mind that the omission of any of these dependencies might be due to data sparsity rather than true independence between variables. Data sparsity may also explain the better performance of the much simpler TAN network, which had many fewer model parameters to estimate from the training corpus.

C. Comparison of Proposed Features

In at least four of the five folds, the running score r_{t-1} was omitted as a parent of two variables through Forward Selection, and so retraining the Forward Selection structure without that variable did not degrade model performance very much - its list-level score correlation and item-level agreement are almost identical with that of the full set of features. This implies, surprisingly, that there was not a sequential dependency between test items (as the running score was intended to capture), but the reason it was omitted might be solely due to the failure of the softmax distribution required by the hypothesized structure. The same was also somewhat true for the TAN network and the SVM classifier - omission of r_{t-1} in either of these did result in significant drops in automatic score correlation, but not in item-level agreement.

Predictably, leaving out the Evidence resulted in very low item-level agreement (as low as 0.000 Kappa for the SVM, which is the lowest possible Kappa). Here the TAN network proved to be remarkably resilient to the omission of this feature subset, with a notably high list-level correlation of 0.739 (significantly greater than the baseline with $p \leq 0.07$) based solely on the Underlying variables alone. Could the TAN network be used to predict a child's reading scores from nothing more than demographics of the child and the difficulty of the test? Perhaps to some degree, but this is definitely more plausible than with the SVM - its zero Kappa and null correlation are due to its classifying every test item as "accept." Apparently the trained SVM's hyperplane was unable to separate the classes when the Evidence was omitted. This speaks to the suitability of a generative model (rather than a discriminant model) for performing this task.

TABLE VII
USING THE FORWARD SELECTION PROCEDURE OUTLINED IN SECTION V-B, THESE ARE THE TOTAL NUMBER OF TIMES EACH OF THE HYPOTHESIZED NETWORK ARCS WAS SELECTED FOR THE FINAL REFINED NETWORK, OVER 5 CROSSVALIDATION TRAINING SETS.

			Children						
					E_t				
					e_t^1	e_t^2	e_t^3	e_t^4	e_t^5
Parents			r_{t-1}	q_t	*ROS*	*NA*	*SP*	*RD*	$\hat{p}$
U_t	u_t^1	*item index*		0	5	5	5	5	0
	u_t^2	*word length*		1	5	5	5	5	0
	u_t^3	*word list*	5	5	5	5	5	5	5
	u_t^4	*L1*	5	5	5	5	5	5	5
	u_t^5	*gender*	5	4	5	5	5	5	5
	u_t^6	*grade level*	5	4	5	5	5	5	5
		r_{t-1}		1	5	5	5	5	0
		q_t			5	5	5	5	5

Omitting the Underlying variables did reduce all 3 classifiers' performance significantly, but without as much of a drop as brought on by omitting the Evidence. This suggests that, in making a reading score, the Evidence variables are far more important than the Underlying ones. However, it is interesting to note that we can predict surprisingly accurate list-level scores merely with prior knowledge of the child and the test, without even observing how the child performs on the test. Furthermore, including the Underlying variables does significantly improve the list-level score correlation.

D. Bias and Disagreement Analysis

Section III reported that listeners gave higher reading scores to native English than Spanish-speaking students, over a small subset of the data (based on a one-tailed test of difference in proportions with $p \leq 0.05$). For Listener #1 (the reference), this bias over the entire corpus extended also to higher scores for females than for male students, and for older students than for younger (2nd grade over 1st, and 1st grade over Kindergarten). This background info was not given to any of the listeners, but we can assume they guessed each child's gender, age, and native language from their voices. A conscientious listener would try to judge objectively regardless, so it's possible that these differences in proportions are simply a mirror of the children's performance in the data. In the automatic scores, these same biases (if they were in fact biases) were retained by all the various methods investigated. However, the two baseline methods did not give significantly higher scores to male or female students, nor to English over Spanish-speaking students (or vice versa). A plausible interpretation is that, in improving over the baseline, the non-baseline classifiers learned to imitate the Listener #1's subjectivity so well that they even replicated that listener's biases.

E. Remaining Questions

A number of lingering questions remain. First, the best results suprass only the bottom end of the inter-listener agreement range. How can we improve this? In the student

model results, we saw a dramatic decline in performance with the exclusion of the Evidence or Underlying features. The Forward Selection procedure on the network structure did not completely exclude any one feature from the model - they all seemed worthwhile in making an automatic reading decision. These findings suggest that the addition of more features - especially along the lines of pronunciation Evidence - can potentially help make the automatic scores more human-like.

What about the automatic elimination of any dependencies that would have required softmax distributions - is there another type of PDF that could be used for those discrete variables that require continuous parents? Should the parent variables in these situations be discretized, or could the children be made continuous without using softmax functions? Seeing as Forward Selection did not eliminate these variables entirely, the ones in question must be valuable to the model. Finding a proper way to represent their probability distributions would be another potential source of improved performance.

VIII. Conclusion

This paper proposed a new student model with a number of unique features for use in automatic scoring of reading skills when demonstrated by isolated words read out loud. First, we explained why this was not simply a pronunciation evaluation or verification task, then we suggested some new features that should be useful in such a model - cues that we expected teachers to use when judging reading ability, like pronunciation Evidence and Underlying information about the child or test. Then we described a hypothesized Bayesian Network structure that would account for the potential conditional dependencies among all these features and reflect the way we expect teachers might conceive of a student's cognitive state when reading. We also proposed a method for automatically refining this network by using a greedy forward-selection of the conditional dependencies.

Our experiments on the TBALL dataset revealed that the use of these novel features instead of simple pronunciation verification can result in a significant increase in agreement and correlation of automatic scores with human perception. This spoke to the usefulness of the proposed features based on various pronunciation lexica that illustrated different reading or accent phenomena, as well as the child demographics. We also found that our network refinement algorithm did not choose to exclude very many of the hypothesized arcs to improve performance, which testifies to the sound basis of the initial proposed network. Our best network models exhibited the same potential biases as the human scores, a testament to how well they had learned the human annotators' trends in making subjective judgments.

The methods presented here may seem specific to reading assessment, but could they be used elsewhere with only some minor modifications? To construct a similar student/cognitive model, one need only specify the Evidence and Underlying features that apply to the given task. Such features can then be united using the network and refinement algorithm presented here, with no real changes necessary. It could potentially be used for other types of assessment and pedagogy (pronunciation training, math tutoring, etc.), or even user modeling for a dialogue system or emotion recognition.

References

[1] M. J. Adams, *Beginning to read: Thinking and learning about print*. Cambridge: the MIT Press, 1990.

[2] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," in *Proc. of MMSP*, Chania, Greece, Oct. 2007.

[3] E. Atwell, P. Howarth, and C. Souter, The ISLE corpus: Italian and German spoken learners English, *ICAME Journal*, vol. 27, pp. 518, 2003.

[4] T. M. Bailey and U. Hahn, "Phoneme similarity and confusability," in *Journal of Memory and Language*, 52: 347-370, 2005.

[5] S. Banerjee, J. E. Beck, and J. Mostow, "Evaluating the Effect of Predicting Oral Reading Miscues," in *Proc. of Eurospeech*, Geneva, 2003.

[6] J. E. Beck and J. Sison, "Using knowledge tracing to measure student reading proficiencies," in *Proc. of the 7th International Conference on Intelligent Tutoring Systems*, Sept. 2004.

[7] C. Conati, A. Gertner, and K. VanLehn, "Using Bayesian Networks to Manage Uncertainty in Student Modeling," in *User Modeling and User-Adapted Interaction*, 12(4):371-417, 2002.

[8] A. T. Corbett, J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," in *User Modeling and User-Adapted Interaction*, 4(4):253-278, 1994.

[9] C. Fought, *Chicano English in Context*. New York: Palgrave MacMillan, 2003.

[10] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," in *Machine Learning*, 29(2-3):131-163, 1997.

[11] C. Goldenberg, "Teaching English Language Learners," in *American Educator*, 32(2):8-21, 2008.

[12] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition With Application to Interactive Books and Tutors," in *Proc. of ASRU*, St. Thomas, 2003.

[13] A. J. Harris and M. D. Jacobson, *Basic Reading Vocabularies*. New York: MacMillan, 1982.

[14] M. Hasegawa-Johnson, J. Cole, K. Chen, L. Partha, A. Juneja, T. Yoon, S. Borys, X. Zhuang, "Prosodic hierarchy as an organizing framework for the sources of context in phone-based and articulatory-feature-based speech recognition," in S. Tseng (Ed.), *Linguistic Patterns of Spontaneous Speech*, special issue of *Language and Linguistics*, Academica Sinica, in press.

[15] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," in *J. Acoust. Soc. Am.*, 121(2):723-42, 2007.

[16] J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane, "A Prototype Reading Coach that Listens," in *Proc. of AAAI-94*, Seattle, 1994.

[17] K. P. Murphy, "The Bayes Net Toolbox for Matlab," *Computing Science and Statistics*, vol. 33, 2001.

[18] K. P. Murphy, "A Variational Approximation for Bayesian Networks with Discrete and Continuous Latent Variables," in *Proc. of the Conf. on Uncertainty in AI*, 1999.

[19] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implication for reading instruction," Tech. Rep. 00-4769, National Institute for Child Health and Human Development, National Institute of Health, Washington, DC, 2000.

[20] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, 30(2-3):83-94,1999.

[21] J. Reye, "Student Modelling based on Belief Networks," in *International Journal of Artificial Intelligence in Education*, 14:1-33, 2004.

[22] J. Shefelbine, *BPST - Beginning Phonics Skills Test*, 1996.

[23] K. Stevens, "Features in Speech Perception and Lexical Access," in *The Handbook of Speech Perception*. Ed. D. B. Pisoni and R. E. Remez. Oxford: Blackwell, 2005.

[24] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, "Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners," in *Proc. ICSLP*, 2000.

[25] J. Tepperman and S. Narayanan, "Using articulatory representations to detect segmental errors in nonnative pronunciation," in *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):8-22, Jan. 2008.

[26] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, "A Bayesian Network Classifier for Word-level Reading Assessment," in *Proc. of InterSpeech ICSLP*, Antwerp, Belgium, August 2007.

[27] J. Tepperman, 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. of InterSpeech ICSLP*, Pittsburgh, 2006.

[28] P. Westwood, *Reading and Learning Difficulties: Approaches to Teaching and Assessment*. Camberwell, Victoria: ACER, 2003.

[29] S.M. Williams, D. Nix, P. Fairweather, "Using Speech Recognition Technology to Enhance Literacy Instruction for Emerging Readers," in *Proc. of Fourth International Conference of the Learning Sciences*, Mahwah, NJ, 2000.

[30] S. Witt and S. Young, "Computer-assisted Pronunciation Teaching based on Automatic Speech Recognition," in *Proc. of the Conference on Language Teaching and Language Technology*, pp. 25-35, April 1997.

[31] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[32] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, Pronunciation variations of Spanish-accented English spoken by young children, in *Proceedings of Eurospeech*, Lisbon, Portugal, 2005.