A Bayesian Network Classifier for Word-level Reading Assessment

Joseph Tepperman[1], *Matthew Black*[1], *Patti Price*[2], *Sungbok Lee*[1], *Abe Kazemzadeh*[1], *Matteo Gerosa*[1], *Margaret Heritage*[3], *Abeer Alwan*[4], *and Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, USC
[2]PPrice Speech and Language Technology
[3]Center for Research on Evaluation, Standards, and Student Testing, UCLA
[4]Speech Processing and Auditory Perception Laboratory, UCLA

Abstract

To automatically assess young children's reading skills as demonstrated by isolated words read aloud, we propose a novel structure for a Bayesian Network classifier. Our network models the generative story among speech recognition-based features, treating pronunciation variants and reading mistakes as distinct but not independent cues to a qualitative perception of reading ability. This Bayesian approach allows us to estimate the probabilistic dependencies among many highly-correlated features, and to calculate soft decision scores based on the posterior probabilities for each class. With all proposed features, the best version of our network outperforms the C4.5 decision tree classifier by 17% and a Naive Bayes classifier by 8%, in terms of correlation with speaker-level reading scores on the Tball data set. This best correlation of 0.92 approaches the expert inter-evaluator correlation, 0.95.

Index Terms: Bayesian network, reading assessment, pronunciation evaluation, children's speech

1. Introduction

To accurately assess a child's reading skills by the pronunciation of the words they read out loud, a teacher must first know to distinguish between simple variants in pronunciation and "true" reading mistakes that betray a lapse in comprehension. Prior knowledge of the child's age, native language, or regional dialect may influence what one believes to constitute an acceptable pronunciation, from a literacy assessment point-of-view [1], and this assessment should be fair regardless of the child's background. Expertise regarding the errors typically made by young readers can similarly affect a teacher's judgment. They may also interpret the present pronunciation in light of its context, or compare it with past readings by this same child, to assess the student's degree of consistency or improvement. All these factors and more serve to inform a teacher's perception of a student's reading ability.

The complexity of the task and the need to incorporate adequate contextual cues accounts for the fact that the majority of past work in the use of automatic speech recognition for tutoring reading has focused on utterance-level assessment [2]. However, for younger children (ages 4-6), eliciting paragraph- or sentence-level reading is not always feasible or informative. Teachers will often measure a young child's reading ability by calculating the speed and accuracy of isolated words read from a list - an important component in early literacy assessment that helps teachers understand how reading skills are developing [3]. The aim of this study was to automatically judge this aspect of a new reader's literacy, in the case of word-level elicitation.

We have previously addressed this task by formulating it as a traditional pronunciation verification problem: given a speech observation, we compare the likelihood that it was drawn from some target model's distribution to the likelihood that it comes from a filler model of expected reading mistakes [4]. Advances along these lines involve creative ways of training or defining the filler model so as to discriminate best between pronunciation classes [5]. Though this method is useful in pronunciation evaluation, it is not always appropriate for reading assessment, which is strictly-speaking not merely a pronunciation evaluation problem. In a more complex machine-learning framework such as a decision tree classifier, the addition of features beyond likelihood scores - including features derived from recognition n-best lists and speaker demographics - has been shown to provide some limited improvement [4]. However, correlations between representative features used in this study are as high as 0.8 or 0.9; this indicates that by simply adding more recognition-based features we should not expect much improvement in machine discrimination between target and filler.

A likelihood ratio threshold or pruned decision tree will normally not consider all available cues in making an automatic reading assessment decision. The class decisions themselves are not definitive so much as they are perceptual and open to dispute on a continuous scale, but a decision tree classifier cannot return a true continuous posterior score for a class given a set of features. And the high inter-feature correlation makes many of the features redundant unless their probabilistic dependencies are trained as model parameters. Here, we argue that these conditions lend themselves to modeling this task under a Bayesian Network framework.

Bayesian Networks have recently been used for many speech-related applications, including pronunciation modeling and speech recognition [6]. Given a discrete class variable Q and a vector of features $X_1, X_2, \ldots, X_n$, a Bayesian Network classifier defines the joint probability

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

where $Pa(X_i)$ - the "parents" of feature X_i - refers to all features which, if known, we would expect to influence the distribution of X_i - all other features we assume to be independent [7]. To use Eqn. (1) as a classifier for Q, every feature X_i must have at least Q as one of its parents (i.e. Q is the network's "root node"), and a classification decision is made by choosing a value for Q that maximizes this expression. In the case that each feature has only Q as a parent, then all features are assumed to be independent - this is the Naive Bayes case. In our study, Q is a binary assessment of word-level reading quality: either acceptable or unacceptable, though the posteriors for each class can be used to calculate softer decision scores.

A correlation between features may or may not betray a causative relationship between them, but if we have reason to believe that one feature in a sense "generates" another, then we can model the former feature as the latter's parent, and knowledge of the parent feature's value will influence our expectations of its child's probability. We propose a new structure for a Bayesian Network classifier that can tease apart the various factors that contribute to perception of word-level reading errors and model the probabilistic dependencies among these highly-correlated features, allowing for a final reading assessment score that reflects all available cues and their interactions.

2. Corpus

The speech recordings used in this study come from data collected in Los Angeles schools as part of the Technology-Based Assessment of Language and Literacy (Tball) project [3]. In a classroom environment with close-talking microphones, children in Kindergarten through Grade 2 were asked to read aloud isolated words elicited by an animated user interface. Our test set's targets were comprised of words appropriate for Kindergarten and Grade 1 students in order of increasing difficulty, and test conditions were as close as possible to those of a school teacher's ordinary literacy assessment of this type.

Because many of the children were of Mexican background, we also collected demographic information regarding each student's native language. Of the 29 students that made up our test set, 11 of them were native English speakers, 11 were nonnative, and for the remaining 7 this statistic was missing, possibly because these children's parents chose not to answer our survey.

3. Choice of Features

Our choice of features reflects the desire to model true reading mistakes and simple pronunciation variants as distinct, though not independent, entities. Both may contribute to a teacher's perception of a child's reading skills, though often the exact relationship is unclear [1]. We began by constructing a lexicon of pronunciation variants spanning four classes: the set of expected variants of the target word for this data set, **TA**; the set of expected reading mistakes and guesses, **RD**; the set of expected variants given the possible influence of nonnative letter-to-sound rules from the child's first language (in this case, Mexican Spanish), **L1**; and a background or silence set to detect when the child is unable to produce the target and offers no response, **SIL**. For example, for the target word "Can," **TA** variants include /kæn/ and /kɛn/, while for an **L1**-influenced pronunciation we would expect /kan/, and a common **RD** mistake would be to make the vowel say its name - /keɪn/. The selection of these pronunciations was based on a previous study in common substitutions made by children in the Tball corpus [8], and with extensive guidance from experts in literacy and phonetics. With our lexicon augmented with tags related to each of these pronunciation classes, we calculated the following features: the likelihood of the observed speech given each class, $P(O|C_i)$, where C_i is defined as an unweighted network of all pronunciations within class i and i can take values $\{$**TA**, **RD**, **L1**, **SIL**$\}$, and four binary features that encode the pronunciation class or classes of the best recognition hypothesis, $\mathrm{argmax}_i\, P(O|C_i)$, accounting for pronunciations common to more than one class.

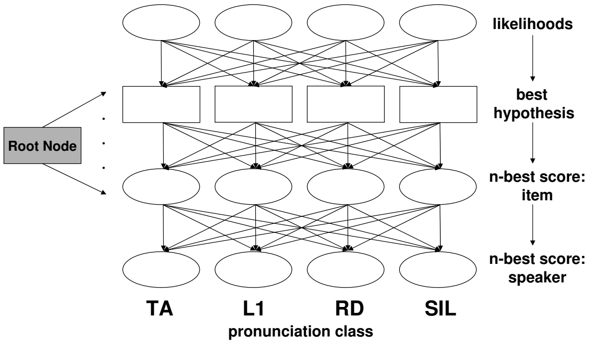

Figure 1: Part of our proposed Bayes Net structure, with four tiers of recognition features parameterized into four pronunciation classes. Rectangles and ovals denote discrete and continuous nodes, respectively.

In the end, recognition of any one of these pronunciations cannot reliably inform a categorical reading assessment, since many of the variants are so close as to be be practically indistinguishable with HMM models. So, in addition to the best recognition hypothesis among these pronunciation classes, a more reliable set of recognition-based features for a target item would be the percentage of tokens in the n-best recognition results that match each of the four pronunciation classes, referred to hereafter as our four "n-best scores." Due to variants in exact frame-level model alignment and the multiple pronunciations within each class, we would expect the same class tag to appear multiple times on a list of recognition hypotheses.

Along this same line of thinking, even if we could reliably discriminate between close pronunciations, we are not always able to declare with certainty that a particular pronunciation always indicates satisfactory or poor reading skills (except in the case of **SIL**). In some cases, reading "Can" as /kɛn/ will indicate satisfactory reading skills, in others it will not, and many factors besides segment-level pronunciation influence this perception of acceptability. A word in isolation is usually not enough evidence to judge a speaker's pronunciation [9], much less assess their literacy. However, an individual item's assessment can be informed by features indicative of the child's pronunciation overall. To this end, for each child we calculated four global n-best scores - one for each of the pronunciation classes - by taking the mean of all single-item n-best scores for that child, $\frac{1}{k}\sum_k n_i(k)$ where $n_i(k)$ is class i's n-best score for this child's kth target word. In a sense the algorithm becomes non-causal at this point, because it can take into account future pronunciations in the present item's assessment. Additional features include the following: the recognized target's estimated duration, as a cue to fluency; the target word's length (in characters) and order in the list, as measures of difficulty; and the child's grade and native language as discrete variables, though the latter is missing for some of the speakers. Further discussion of some of these features can be found in [4].

4. Network Structure

Though the Naive Bayes classifier has been shown to perform well under many experimental conditions [7], modeling probabilistic dependencies among the features can improve performance. However, finding the best network structure for a given set of features can be a computationally intractable problem, because it requires comparing all directed acyclic graphs for which Q is the root. One proposed solution is the Tree-Augmented Naive Bayesian (or TAN) algorithm, which restricts the search such that each feature is allowed only one parent besides the root node [7], but this did not seem appropriate here.

In terms of a "generative story" that unites our features derived from speech recognition results, we conceive of them distributed over four cascaded "tiers" representative of the steps in the feature extraction procedure, each tier parameterized into the four pronunciation classes outlined in Section 3. Initially, likelihood scores across each class are compared to select the best hypothesis from among those classes - a first tier of likelihoods clearly can be modeled as parents of a second tier of best hypothesis features. Then, a third tier comprised of n-best scores is calculated from all available recognition hypotheses, including the best one encoded in the second tier - again there is clearly a generative relationship from the second to third tiers. Finally, a fourth tier of speaker-level n-best scores can be modeled as children of the third tier's item n-best scores, since it is the third tier's features that contribute to calculating the fourth tier. A graphical depiction of this part of our proposed network is shown in Figure 1. Additional generative relationships were made by modeling the item's duration as another parent of the second tier, the target word length as a parent of the **RD** item n-best score, and the child's grade and native language as predictors of speaker-level **RD** and **L1** n-best scores, respectively.

One evident drawback of this structure is, with the many child-parent relationships hypothesized, a large amount of training data may be necessary to reliably estimate the number of model parameters. Because our model uses both continuous and discrete features in all parent/child combinations, we chose to model many of our discrete features - in particular the second tier described above - as multinomial logistic (or softmax) distributions [10]. In a Bayesian network, a discrete node with discrete parents is typically parameterized as a table of conditional probabilities over all combinations of the parents. For the case of discrete nodes with continuous parents, one option is to artificially discretize all the features, though this usually results in poor parameter estimates and is sometimes computationally unfeasible. If we instead model these nodes as a continuous multinomial logistic distribution, it will behave like a soft decision threshold. In the absence of a large amount of training data, this type of distribution can avoid defining discrete nodes in terms of possibly inaccurate conditional probability tables. Additionally, this distribution requires iterative estimation of all model parameters (EM training), which can also help offset a relatively small number of training instances or the case of missing features, both of which apply in this study.

5. Perceptual Evaluations

For training and testing the Bayesian Network classifier, the set of 29 students mentioned in Section 2 were judged acceptable or unacceptable on the item level by one expert listener - 442 target words in all, averaging about 15 per speaker. To assess inter-evaluator agreement as an upper-bound for classifier performance, a different set of Tball recordings from 13 speakers were judged acceptable or unacceptable on the item level by 14 listeners who rated all items from each child. We also calculated an overall score for each child/listener pair as the percentage of items judged acceptable for that speaker. Agreement and correlation results, sorted by type of evaluator, are given in Table 1.

We found that, though the teachers had better item-level Kappa agreement than the non-teachers, both groups had high inter-evaluator correlation in the percentage-based speaker-level scores, which we argue are more important than item-level assessments, and this result indicates that the percentage of items judged as acceptable is a reliable indicator of overall speaker-level reading ability. Many studies such as in [9] have asserted the ambiguity of pronunciation perception when based on isolated words, and we would not necessarily expect an automatic classifier to make robust item-level judgments for reasons stated in Section 3.

	teachers	*non-teachers*	*all*
# of evaluators	5	9	14
Kappa agreement	0.847	0.753	0.788
corr.: % acceptable	0.951	0.923	0.934

Table 1: Mean inter-evaluator agreement and correlation as judged by teachers and non-teachers.

6. Experiments

With about 19 hours of classroom recordings taken from both native and nonnative speakers, we trained context-dependent phone models from this corpus. Annotated on the word level, we used canonical pronunciation expansions and the Baum-Welch algorithm to estimate three-state HMM parameters with 16 Gaussian mixtures per state. Generic models for the background classroom noise were trained by cutting background segments out of the recordings and estimating their parameters separately - here 256 mixtures per state proved necessary for adequate target word endpointing performance.

From the set of 29 speakers described in Section 2, we extracted all item- and speaker-level features explained in Section 3. All n-best scores were calculated with n equal to 20. This set was partitioned by speaker into ten folds for cross-validation. For purposes of comparing our Bayesian Network classifier's performance with that of others, we trained four different classifiers over eleven feature sets. The four classifiers we chose were the C4.5 decision tree, a Naive Bayes classifier, our fully-connected Bayesian Network described in Section 4, and a refined version of our network based on feature selection (explained below). The eleven feature sets were defined by removing different categories of features from the complete set (1): each of the four tiers (2-5), each of the four pronunciation classes (6-9), the item information features (10), and the demographics features (11). For the network classifiers, when one tier of features was omitted, the tiers immediately preceding and succeeding it were connected. All Bayesian networks were implemented in BNT [10], and the decision tree was trained with the Weka toolkit [11]. Though these classifiers operated on the item level, we calculated continuous speaker-level scores as the mean of all item-level posteriors for the "acceptable" class.

The refined version of our Bayesian Network was constructed by disconnecting from the root node the subsets of features which degraded the performance of the Naive Bayes and C4.5 classifiers in terms of Kappa agreement and speaker-level correlation. These features, though disconnected from the root, would still be connected to the other features as outlined in Section 4 and would exert an influence on the final classification decision.

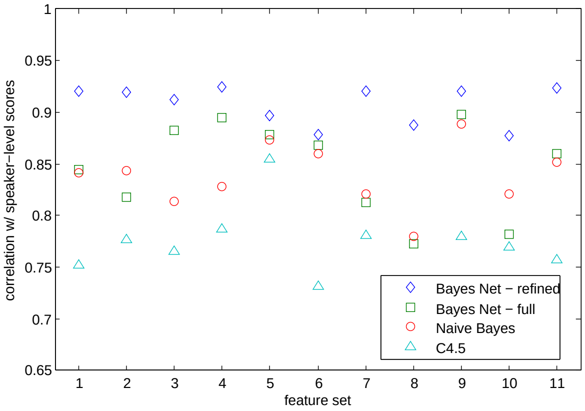

Figure 2: Comparative classifier performance over 11 feature sets, with respect to correlation with speaker-level scores.

7. Results and Discussion

Figure 2 illustrates the relative performances of the four classifiers over the eleven feature sets described in Section 6. The refined Bayesian network outperformed the three other classifiers for all feature sets, while our unrefined network outperformed the remaining two in the majority of cases. Table 2 reports numerical performance of all four classifiers for the set of all features. Here overall speaker-level correlation between automatic and human judgments approaches that of the inter-evaluator agreement reported in Section 5.

The only subsets which never degraded Kappa agreement or correlation when omitted were the demographics (set 11) and those from the **SIL** pronunciation class (set 9), so these subsets were the ones disconnected from the root node in our refined structure. This indicates that perhaps the **SIL** features offered only redundant information when used alongside those of the other pronunciation classes, and that our chosen child demographics are not reliable indicators of reading ability (though this may be due to the fact that some of these features were missing).

In analyzing the item-level errors made by this best version of our classifier, we found that the difference in proportions of classifier errors for native and nonnative speakers was not statistically significant on the 95% confidence level. This suggests that our classifier is not biased against either category of student. However, there were statistically significant differences in the proportions of errors seen between first graders and second graders - suggesting that age-dependent modeling might be necessary. There was also a significantly lower proportion of classifier errors on shorter target words (one or two letters long) in comparison with longer target words (three or four letters long), probably due to the growth in the number of expected pronunciation variants as the length of the target word increases.

8. Conclusion

In conclusion, with our proposed network structure we could assess a speaker's reading ability with correlation approaching that of expert inter-evaluator agreement. This new network consistently outperformed the state-of-the-art C4.5 decision tree classifier, and with all features a refined version of the network had 8% improvement in correlation compared to a Naive Bayes approach. Though the network structure is somewhat complex, with EM training and the use of softmax nodes to model some discrete variables, we found that we could achieve these performance improvements even with a relatively small amount of training data. This new structure for modeling the generative story among recognition-based features can be adapted for use outside the domain of literacy assessment.

	Kappa agreement	*obj. score corr.*
C4.5	0.535	0.752
Naive Bayes	0.617	0.841
Full Bayes Net	0.641	0.844
Refined Bayes Net	0.681	0.921

Table 2: Comparative performance of our refined Bayes Net classifier on the set of all features (set 1 in Figure 2).

9. Acknowledgements

This work is supported in part by IERI award number 0326214 from the National Science Foundation.

10. References

[1] W. Labov and B. Baker, "What is a reading error?" http://www.ling.upenn.edu/ wlabov/Papers/WRE.html

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

[3] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Anderson, S. Narayanan, and A. Alwan, "Tball Data Collection: the Making of a Young Children's Speech Corpus," in *Proc. of Eurospeech*, Lisbon, 2005.

[4] J. Tepperman, J. Silva, A. Kazemzadeh, H. You, S. Lee, A. Alwan, S. Narayanan, "Pronunciation Verification of Children's Speech for Automatic Literacy Assessment," in *Proc. of Interspeech ICSLP*, Pittsburgh, 2006.

[5] D. Willett, A. Worm, C. Neukirchen, and G. Rigoll, "Confidence Measures for HMM-Based Speech Recognition," in *Proc. of ICSLP*, Sydney, 1998.

[6] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," in *Proc. HLT/NAACL*, Boston, 2004.

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

[8] H. You, A. Alwan, A. Kazemzadeh, and S. Narayanan, "Pronunciation Variations of Spanish-accented English Spoken by Young Children," in *Proc. of Eurospeech*, Lisbon, 2005.

[9] L. M. Arslan and J. H. L. Hansen, "Language Accent Classification in American English," *Speech Communications*, vol. 18(4), pp. 353-367, July 1996.

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

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