STAR: Articulation Training for Young Children

H.Timothy Bunnell, Debra M. Yarrington, and James B. Polikoff

The Alfred I. duPont Hospital for Children
and University of Delaware

ABSTRACT

The Speech Training, Assessment, and Remediation (STAR) system is intended to assist Speech and Language Pathologists in treating children with articulation problems. The system is embedded in an interactive video game that is set in a spaceship and involves teaching aliens to "understand" selected words by spoken example. The sequence of events leads children through a series of successively more difficult speech production tasks, beginning with CV syllables and progressing to words/phrases. Word selection is further tailored to emphasize the contrastive nature of phonemes by the use of minimal pairs (e.g., run/won) in production sets. To assess children's speech, a discrete hidden Markov model recognition engine is used[1]. Phone models were trained on the CMU Kids database[2]. Performance of the HMM recognizer was compared to perceptual ratings of speech recorded from children who substitute /w/ for /r/. The difference in log likelihood between /r/ and /w/ models correlates well with perceptual ratings of utterances containing substitution *errors*, but very poorly for correctly articulated examples. The poor correlation between perceptual and machine ratings for correctly articulated utterances may be due to very restricted variance in the perceptual data for those utterances.

1. INTRODUCTION

According to the National Information Center for Children and Youth with Disabilities there were more than one million students in American public schools classified as having a speech or language disorder in 1994 and 1995. Historically, the largest subgroup of children with speech or language disorders are those with phonological disorders[3]. Most of these children can benefit from speech training interventions, but clinics are often overtaxed for service delivery causing delays in starting children in therapy and reducing the frequency with which children can be seen. Access to therapeutic services would be enhanced by relieving some aspects of speech therapy from Speech Language Pathologists (SLPs) through the use of computer-aided training procedures.

The present work is intended to address the need for expanding access to speech therapeutic services by providing software for speech evaluation and articulation training. The Speech Training, Assessment, and Remediation (STAR) system would not replace SLPs, but would facilitate their assessment of speech by helping them to better target therapeutic intervention, augment their efforts in highly repetitive articulation drill and training, and assist in record keeping and reporting. A child will be able to use the software on a home computer, interacting with the software via an animated computer character. Since the system will be constantly eliciting speech from a child and measuring the speech produced by the child, it will be capable of extensive record keeping and report generation, further assisting clinical staff in their duties.

Computer assisted speech training methods have been studied extensively since the early 1980's, and fall into roughly three categories: systems that present speech, and may elicit speech, but do not attempt to analyze or assess speech; systems that provide biofeedback of speech-related acoustic or physiological measures such as amplitude or fundamental frequency; and systems that provide assessment of speech and feedback on the accuracy of speech production. Discussion is restricted here to systems of the latter type that use only the acoustic speech signal as input.

A number of laboratory systems have been described which employ speech recognition (SR) technology in speech and/or language training[4][5][6][7][8][9]. Two of these systems ([5][6]) are specifically designed to recognize children's speech. The LISTEN system [5] is intended for literacy intervention in children who are at risk for reading delays.

The system described in [6] is intended as an articulation training aid for children speaking British English. The system compares a child's utterance of a known word to a continuous HMM of the word, and also to a model of general speech formed by pooling phone HMMs. If the word-specific model is found to provide a better fit to the utterance than the general speech model, it is assumed that the utterance is a good exemplar of the intended word, otherwise, it is assumed to be a poor model of the word. In laboratory tests, this system achieved about a 30% error rate in distinguishing between correct and minimally contrasting incorrect words (e.g., *bit* versus *pit*) when false alarm and miss rates were equal. Although the HMMs were phone based, the speech assessment of the system was word based.

IBM's commercial speech training product and at least one experimental system [4], use recorded templates of the child's best production as standards against which to measure acceptability of new utterances. Tests of the system in [4] estimate that a recognition error rate as high as 20% is acceptable for articulation training. That error rate is within the capabilities of a small vocabulary speaker dependent system. Speaker dependent isolated word systems have the advantages of being (a) well-tuned to the individual speaker, and (b) afford clear acoustic goodness of fit measures for assessing the speech. Unfortunately, these advantages are vitiated if the segment to be trained is not stimulable for the child. Such systems have the further disadvantage of potentially cumbersome training procedures before they can be used by a child.

A number of systems intended for training foreign language pronunciation provide alternatives to the template matching strategy that might also be useful for evaluating articulation errors[7][8][9]. These systems use HMMs to estimate phone-level segmentation of known utterances and goodness of pronunciation measures based on segment level statistics derived from likelihood or posterior probability. These measures compare an individual's production of the phone sequence to the population of talkers used to train the HMM and are speaker independent. Such systems could be used to assess a child's articulation errors whether or not it is possible to elicit a good exemplar from the child as a template.

Generally, HMM-based systems provide pronunciation assessments that correlate well with human assessment of the same speech at the talker level, that is, averaged over multiple sentences spoken by a given talker[7][9]. However, the correlation is weaker at the individual sentence level, that is, when the HMM assessment of a single sentence is compared to the human assessment of the same sentence. Less work has been published regarding the correlation between HMM assessment and human listener assessment at the individual phone level although this issue is addressed in [8]. Unfortunately, it is this latter measure that is of most interest to speech training applications.

In the following, we examine the performance of the SR engine used in the STAR system in assessing pronunciation of the segment /r/ in children who tend to substitute /w/, or a /w/-like segment for /r/. Performance of the SR assessment is compared to that of human listeners in rating the quality of individual word initial /r/ tokens.

To evaluate the effectiveness of the HMM recognizer for rating the accuracy and/or quality of children's productions, utterances were collected from children being seen for speech therapy in the duPont Hospital for Children.

2. METHOD

There were three components to the evaluation study. First, utterances were recorded from children in the clinic. Second, a set of 56 utterances, all intended to be productions of the word *Rhonda* (the name of a screen character in the game), were selected for perceptual evaluation by students in an undergraduate Speech Pathology course at the University of Maryland.[1] Finally, comparisons were made between the perceptual ratings and segment level measures based on the recognition engine. Each of these is described more fully below.

2.1. Speech Data Collection

Subjects. Eight children between the ages of four and seven who were receiving speech therapy were recruited to provide recorded speech examples. All children were being seen because they were substituting /w/ for /r/ and /t/ for /k/ word initially. Most children had other articulation problems as well.

Speech Stimuli. The words *Rhonda, Wanda, Coco,* and, *Toto* were spoken in isolation by the children. These words were the names of characters in the video game.

Procedure. The children were shown a mock-up of the STAR system video game that ran on a Toshiba notebook computer under the control of the Speech Language Pathologist (SLP). The SLP was instructed to use the mock-up game during therapy to elicit productions of the four target words approximately 10 times each. The SLP was given complete discretion on when during the therapy session to collect the data and on the order in which the stimuli were prompted. The mock-up system logged the word requested by the SLP (as signaled by a keystroke) and recorded the child's utterance directly to disk using 16-bit PCM format at a 22050 kHz sampling rate.

2.2. Perceptual Data Collection

Stimuli. Recordings from the clinic were screened by laboratory staff to select as many examples as possible in which the child was asked to say *Rhonda*. One child was asked to record only one instance of *Rhonda* and one or more examples for each of the other children had to be discarded because (a) the SLP was speaking simultaneously with the child, (b) a loud extraneous noise occurred simultaneously with the utterance, or (c) no utterance was recorded. In all, 56 usable examples were found for use in the perceptual study.

Listeners. The listeners were 50 undergraduate students in the Speech and Hearing Department at the University of Maryland. Twenty-five of these students were in an undergraduate speech pathology class, and an additional 25 were in an undergraduate acoustic phonetics class. There were no students in common for the two classes.

Procedure. Four independently randomized lists of the 56 stimuli were prepared and recorded to audio CD. Each stimulus interval on the CD consisted of the synthetically spoken sequence number of the stimulus (i.e., a number from one to 56) followed by 500 msec of silence followed by the stimulus. There was a three second interstimulus interval between the end of one stimulus and the onset of the spoken sequence number for the next trial. Students sat in their normal classroom locations and heard the stimuli presented over loudspeakers. Each student was asked to rate the quality of the initial /r/ of each stimulus item on a scale ranging from 1 (poor) to 5 (excellent) using answer sheets that were provided. Each class of 25 students heard two of the four randomized presentation lists.

2.3. HMM assessment

Stimuli. The 56 utterances used for the perception study were used for this study as well.

HMM Training. The SR engine was adapted from the DHMM system described in [1]. The 62 context independent phone models vary from one to 14 states depending upon the average phone duration measured in the TIMIT database. There are three silence models: three-state utterance initial and utterance final silence models and a one-state, jump-able, utterance medial silence model.

The models were initially trained on speech from the CMU Kids

Corpus[2]. For the training, the Summer `95 subset was used (these were the "good readers"). The labels supplied with the CMU database used a different symbol set than that commonly used in the SRL. Consequently, the CMU phonetic symbols were first converted to SRL phonetic symbols, and then all sentences in the Summer `95 dataset were relabeled with SRL phonetic symbols using the SRL alignment tool. This tool is based on adult models trained from the TIMIT database, but nonetheless aligned segment labels with greater accuracy than the alignments supplied with the database. The dataset so labeled was then used to train new models and the segment alignment iteratively adjusted.

Procedure. The HMM phoneme models for the sequence /#randx#/ were aligned to the acoustic feature vectors to obtain both the global log likelihood (LL) for the utterance and the local LL associated with the sequence of vectors aligned to the states associated with /r/ (LLr). Several likelihood ratios were then computed to form the basis of SR ratings of /r/ quality. These ratios were determined by fitting alternative phone models in place of the /r/ segment to determine corresponding LLx measures and then estimating /r/ quality as Qr = LLx – LLr. The alternative segments examined were /w/, the syllable final /r/ allophone, the rhotacized schwa, the low back vowel /a/, the syllable initial allophone of /l/, and the palatal approximant /j/.

Since children in this study substituted /w/ or a /w/-like segment for /r/, /w/ was an obvious choice for an alternative. The additional alternative segments were examined on the possibility that they might provide acoustic information not redundant with /w/ that would help in estimating human ratings. This possibility was examined through multiple regression modeling.

3. RESULTS

3.1. Perceptual Data

Figure 1 shows the distribution of quality ratings (averaged over listeners) among the 56 utterances. Responses tended to be very categorical, either a stimulus was rated an excellent /r/ or a fairly poor /r/ (average rating of 2.5 or lower). None of the 56 utterances received average rating in the middle of the scale.

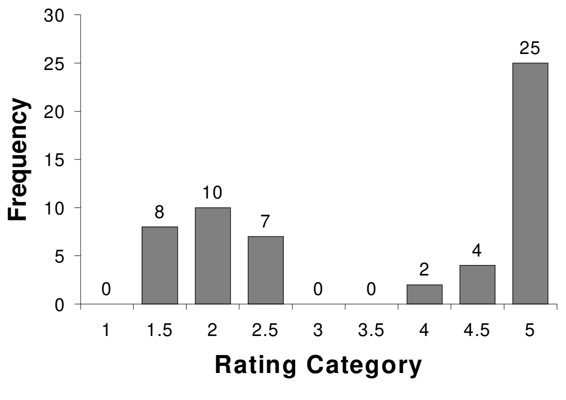

Figure 1. Histogram of average /r/ quality ratings for the 56 utterances. The scale ranged from five *EXCELLENT /r/*, to one *POOR /r/*.

3.2. HMM ratings vs. Perceptual ratings

Figure 2 shows the distribution of likelihood ratios obtained as LLRwr = LLw – LLr for the 56 utterances. These scores are

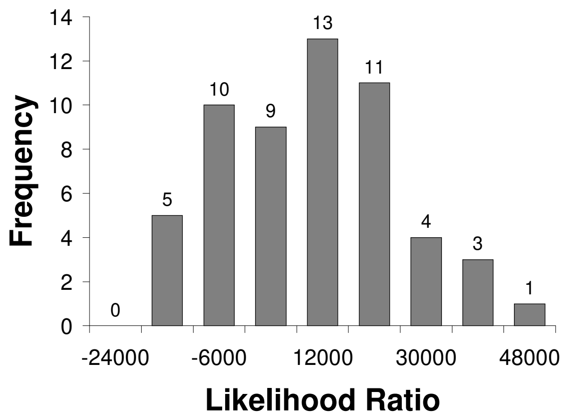

Figure 2 Histogram of HMM based likelihood ratios for the 56 utterances.

clearly more normally distributed than the human ratings.

Nonetheless, regression using only information from /r/ and /w/ models (the /w/ to /r/ phone level LL ratio and the global LL for the /w/ model) provided a significant fit to the human rating data ($F[2,53]=34.32$, $p<0.001$, $r^2= .56$). However, significant further variance was accounted for by including the /l/ to /r/ LL ratio ($t=2.159$ for the added term). This resulted in a model accounting for about 60% of the variance in the human rating scores. The predicted ratings based on the SR data are plotted against the ratings obtained from human listeners in Figure 3. Examination of this figure suggests that much of the variance accounted for may be due to moving the centers of gravity of

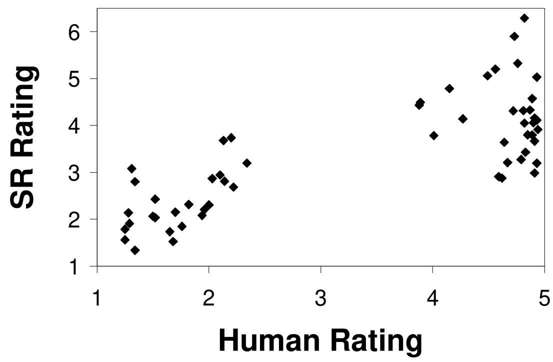

Figure 3 SR ratings versus human ratings of /r/ quality from the best fitting regression model for all utterances.

the two clouds of data, however, there is possibly a linear trend among the data points corresponding to utterances that received poor quality ratings from human listeners.

To explore this possibility further, the data were separated into two classes: those which received a human rating > 3.0 and those which did not. Separate regression models were then fit to these two subsets of the data. No combination of variables provided a significant fit to the data for the good /r/ utterances, however, data for the poor /r/ utterances were predicted strongly by likelihood ratios. The best of these equations in which all terms contributed significantly to the fit was obtained for the combination of LLw and the likelihood ratio LLer = LLer – LLr. These data are plotted in Figure 4 which shows the predicted SR ratings versus the human ratings for only the poor /r/ utterances.

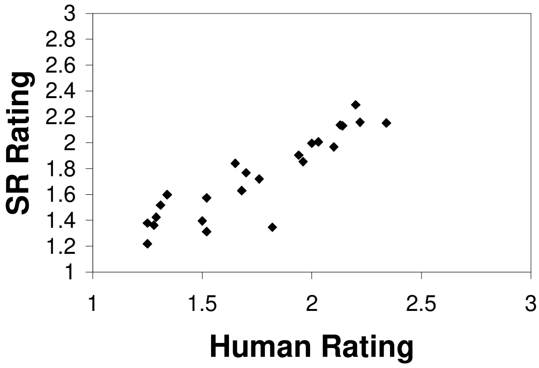

Figure 4 SR ratings versus human ratings of /r/ quality for utterances judged to be poor /r/ exemplars. The SR rating are from the best fitting regression model.

4. DISCUSSION

Scores based on statistics derived from fitting HMM phone models to speech have been shown to provide reasonably good measures of pronunciation accuracy for speakers learning a foreign language. The best correlations between SR-based pronunciation assessment and that of human judges have been obtained when global assessments of speakers are compared by pooling results over multiple utterances. Generally, the more utterances pooled, the better the correlation. Weaker correlations between SR and Human raters are obtained when ratings of individual utterances are compared, and very little work has examined correlations between SR raters and human raters for individual segments within utterances. This latter comparison, however, is crucial to evaluating the effectiveness of SR for articulation training wherein the adequacy of an individual segment in an utterance is the primary factor of interest.

In the present study, SR ratings of /r/ quality were obtained by fitting linear regression models using segment-level likelihood ratios to predict human ratings. For the 56 utterances in this preliminary dataset the results are reasonably encouraging. Overall, one regression model was found to account for about 60% of the variance in the human ratings, and a slightly different model accounted for about 80% of the variance in the human rating data for the subset of utterances that were heard as relatively poor exemplars of /r/. Part of the reason for the absence of a predictive model for the good exemplars of /r/ may be the restricted variance in the human rating data for utterances in this category. By contrast, ratings of utterances in the poor /r/ category were distributed more broadly over a range of scores. Future work will examine whether regression models of the sort applied here can be found to generalize to larger numbers of talkers and utterances and unseen data.

5. REFERENCES

1. Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., and Bunnell, H.T. (1996). The Nemours Database of Dysarthric Speech, Proceedings of the Fourth International Conference on Spoken Language Processing, October 3–6, Philadelphia, PA, USA.

2. Eskenazi, M. and Mostow, J. (1997). The CMU KIDS Speech Corpus. Corpus of children's read speech digitized and transcribed on two CD-ROMs, with assistance from Multicom Research and David Graff. Published by the Linguistic Data Consortium, University of Pennsylvania.

3. Weiner, F. F. (1981). Treatment of phonological disability using the method of meaningful minimal contrast: Two case studies. *Journal of Speech and Hearing Disorders*, 46, 97-103.

4. Kewley-Port, D., Watson, C.S., Elbert, M., Maki, D., and Reed, D. (1991). The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies. *Clinical Linguistics and Phonetics*, 5, 1, 13-38.

5. Mostow, J. (1994). A prototype reading coach that listens. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94),* American Association for Artificial Intelligence, Seattle, WA, August 1994, 785-792.

6. Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., and Barker, P. Applications of automatic speech recognition to speech and language development in young children, *ICSLP 96 – Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, USA

7. Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30, 2-3, 83-94.

8. Witt, S.M. and Young, S.J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning, *Speech Communication*, 30, 2-3, 95-108.

9. Cucchiarini, C., Stirk, H., and Boves, L. (2000). Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30, 2-3, 109-120.

6. Acknowledgements

This work is supported by the Nemours research programs. The authors are grateful for the assistance of Nan Bernstein Ratner in collecting the human perceptual rating data.