ROBUST RECOGNITION AND ASSESSMENT OF NON-NATIVE SPEECH VARIABILITY

Joseph Tepperman, Jorge Silva, Abhinav Sethy, and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory, University of Southern California

ABSTRACT

Motivated by a desire to assess speaking and reading skills and perform unsupervised tutoring of non-native speakers in a foreign language, robust evaluation of speech variability and pronunciation quality must incorporate perceptually meaningful information from many domains of speech analysis – spectral and prosodic, segmental and suprasegmental, and so on. In this paper we present three techniques for pronunciation evaluation on multiple time scales, as well as details of two example language-learning applications currently being implemented with these methods.

1. INTRODUCTION

Correct phone- and syllable-level pronunciation is crucial for effective communication in any language, and therefore should be of high significance to students of a foreign language. Systematic phone insertion, deletion, or substitution, as well as misplaced stress can potentially alter the perceived meaning of what was spoken, or at the very least hinder a speaker's intelligibility in her new tongue. If such a student is to make any use of a language learning system designed to automatically assess her pronunciation, then that system must be capable of identifying and correcting pronunciation mistakes on the phone level at least as well as a human tutor would.

After being trained with expert transcriptions, the methods outlined in this paper require no human supervision. Since these are designed with language learning modules in mind, in which registered users' pronunciations are evaluated based on utterances spoken after machine prompts, it's safe to assume that the aligner has prior knowledge of said prompts (and their expected transcriptions and most common mistakes), as well as perhaps some modicum of background meta-information on the speaker with which to optimize the evaluation.

2. HIDDEN-ARTICULATOR MARKOV MODELS

Oftentimes an easily predictable or commonly recurring mispronunciation may be a function of the phonological rules and phonetic structure of the speaker's native language [1]. For example, native speakers of German learning English as a second language are likely to substitute /s/ for /z/ in such words as "dessert" or "warnings," because in German the character 's' is often pronounced unvoiced as the phoneme /s/ in these contexts. With a model of speech as a time-series of articulatory events – asynchronous jaw, tongue, lip, velum, and vocal cord movements – /s/ and /z/ represent identical physical configurations, though differ in voicing [2]. For purposes of evaluating non-native speech, articulatory feature models would seem ideal for localizing this type of close error in the physiological domain and providing useful feedback to a second-language student in the realm of speech production. And prior knowledge of these expected, recurring mispronunciations can make a language-learning tool more effective in targeting the most important and difficult errors a non-native speaker might make.

The data used in these experiments was compiled by the University of Leeds in their ISLE corpus [1]. These recordings consist of 46 adult Intermediate British English learners who are native speakers of either Italian or German – 23 of each. Utterance prompts were complete sentences designed to highlight specific difficulties English learners typically encounter in pronouncing single phone pairs, phone clusters, and primary stress pairs. The recordings were automatically segmented by a forced-aligner, then these transcriptions were augmented on the phone level by a team of five linguists to reflect each speaker's pronunciation (though, to keep the training procedure as unsupervised as possible, we did not correct any discrepancies in the automatic segmentation times).

2.1. Choice of Models

Chosen because of their representation in concrete physical terms, the models used in this study are based primarily on the Hidden-Articulator Markov Models proposed in [2], but with some important differences. We built a separate Hidden-Articulator Markov Model for each of the eight features (Jaw, Lip Separation, etc.), and classified every one separately. This allowed for a degree of asynchrony among the articulators, so that the results might mimic the overlapping behavior of a true vocal tract's constituent parts.

These simplifications rest on the assumption of independent motion among these eight articulator streams, which in a different study might not be valid – it allows for results that could potentially violate the fundamental physical constraints of the human vocal tract (e.g. dependencies between the jaw position and lip separation, tongue tip and tongue body, etc.). But the point of this project was not to build an articulator-based speech recognizer or even a general phoneme recognizer. Rather, we intended to demonstrate clustering of correct and incorrect pronunciations in articulatory feature space, regardless of the accuracy in recognizing an individual articulatory feature. In fact, such clustering should perform better under an assumption of independence, since disallowing physically impossible articulatory configurations will seriously limit the representation of fine pronunciation distinctions within articulatory feature

German				Italian			
phoneme - error	*precision*	*recall*	*accuracy*	*phoneme - error*	*precision*	*recall*	*accuracy*
/z/ - /s/	67.59	79.32	57.46	/uh/ - /uw/	74.34	72.51	57.91
/ax/ - /uh/	72.76	78.98	61.04	/ih/ - /iy/	72.18	72.37	56.40
/v/ - /f/	74.40	80.70	63.24	/ah/ - /ax/	79.67	83.71	68.98
/w/ - /v/	76.97	82.74	66.21	/ax/ - /oh/	69.52	75.69	56.81
/uw/ - /uh/	80.57	77.07	64.49	/t/ - /t/ + /ax/	71.70	73.02	56.59
/ah/ - /ax/	72.38	78.68	60.47	/ng/ - /ng/ + /g/	76.38	72.39	59.14
/t/ - /deletion/	71.48	76.96	58.42	/er/ - /eh/ + /r/	70.88	75.39	57.64

Table 1. Percentage results by phoneme, each averaged over three random training and test set partitions.

space. If the results point toward a physically unlikely or noncanonical articulation, it probably signifies the presence of a pronunciation error, and that is exactly what we intend to detect.

2.2. Experiments and Results

With these eight articulatory models and a bigram "language model" to constrain the physical possibilities of articulatory transitions within a given feature, we performed Viterbi decoding for articulation recognition on the same 46 ISLE speakers (a total of about 90 complete sentences per speaker).

[1] lists the most difficult English phones for the German and Italian speakers in the ISLE corpus, along with their associated most common mispronunciations. Starting with the forced-alignment segmentation of the ISLE data, we matched up occurrences of the difficult phones (pronounced correctly or not) to the eight streams of our recognition results. Since forced-alignment segmentation times (provided along with the ISLE recordings, and purposely left uncorrected by the annotators) are sometimes erroneous on the phone level, and since the decoding results would probably overlap asynchronously with the forced alignment times, we executed a "soft" decision scheme that averaged all articulatory results within the alignment segmentation interval, allowing for resulting vectors lying "between" the previously-defined quantization positions. This was done by initially representing each class as an integer number, in a sequence consistent with the physical progression of classes within a given feature, similar to what was done in [2]. For example, the four Jaw Position classes were assigned: 0, Nearly Closed; 1, Neutral; 2, Slightly Lowered; 3, Lowered.

We assigned each occurrence to one of two classes for supervised clustering based on the ISLE transcriptions: correct pronunciation, or its most common expected error. We used these eight-dimensional "soft" decision vectors (one for each occurrence) to generate a separate binary nearest-means classifier for each difficult phone and its corresponding error, empirically deriving the best relative sizes of a random partitioning of the vectors into training and test sets. Performance of this final classification procedure is reported in Table 1.

[1] reported an inter-annotator agreement of "at best" 70% when simply detecting the location of a pronunciation error (but not deciding what the error is). These results all come from unsupervised classification of human-tagged speech, so results around 70% indicate that our method performs as well as a human annotator would. Since this method is intended to evaluate pronunciation in students of a foreign language, a small number of false alarms is tolerable, as that will only err on the side of requiring the student to practice her pronunciation more.

Of course, [1]'s results were averaged over all possible phonemes and errors, and for each native language we only considered seven, with each one's most common error. But the phonemes and systematic errors investigated here account for roughly 20-30% of all phone-level pronunciation errors in the ISLE corpus. Moreover, they encompass different types of mispronunciations on various articulatory levels, with consistent results. Canonical articulations for /ax/ and /uh/ differ in the Jaw, Lip Width, and Tongue Body features, whereas /uw/ and /uh/ differ only in terms of Lip Separation and Lip Width, /v/ and /f/ differ only in Voicing, and so on. So, our results can be thought of as representative of this method's potential performance on the corpus at large.

Considering the relative sparseness of the incorrectly-pronounced data, the results for detection of each of these phoneme's most common errors are consistent between the two native languages and among all phonemes. Even /ah/ and /ax/, though in training mapped to the very same articulatory configuration, could be distinguished with performance comparable to the other, more dramatic misarticulations. But the results in Table 1 suggest that the method proposed here, to be useful in language-learning contexts, is not dependent on any particular native language.

What about other phone-level errors besides the most common unique mispronunciation? In the native German section of the ISLE corpus, /uh/ is substituted for /ax/ almost as often as /oh/, /uw/, /ae/, and /eh/. In distinguishing between a correctly pronounced /ax/ and any error at all, we found the results to be the same as in distinguishing between only /ax/ and /uh/. So this method has the potential to be incorporated into more general pronunciation evaluation tasks, or even phoneme recognition. Given robust models of canonical and non-native phones, articulatory features could be applied to any generic pronunciation evaluation task, regardless of the languages or mispronunciations involved.

3. ANALYSIS OF SEGMENTAL SCORES

This section presents a comparative evaluation of standard HMM segmental-based scores in the task of predicting human judgment about pronunciation quality of non-native speakers. Segmental scores were generated across different types of acoustic models and they were evaluated with respect to

correlation with human judgment of pronunciation quality. In addition, a discriminative acoustic analysis is conducted in the vocabulary of the task, to see if acoustic discrimination measures between native and non-native target word models affect performance of automatic pronunciation evaluation systems.

Segmental scores are generally based on the use of target acoustic models (HMMs), capturing the canonical pronunciation of a given target item, and in addition a sub-family of those scores consider a background acoustic unit, representing all unacceptable acoustic variations of the target unit [3]. The likelihood that a given time-series of observations is consistent with a target model is given by the probability expression $P(O \mid \lambda_i)$. To generate a more refined score of the confidence that O belongs to class λ_i, the classic method [3] is to take a ratio of these likelihood probabilities:

$$\tau = \frac{P(O \mid \lambda_i)}{P(O \mid \lambda_f)}$$

where λ_f is a generalized "filler" (or "background") model for all miscellaneous speech. Taking the log of τ we can turn the likelihood ratio into a difference of log-likelihoods. This final score is considered to be proportional to the pronunciation quality of the acoustic realization [3,4]. Choosing an appropriate threshold T we can empirically optimize a binary pronunciation verification decision: for $\tau \geq T$ we accept, for $\tau < T$ we reject.

The speech material used for this analysis consists of adult native and non-native English speakers producing 12 isolated target words. The corpus contains approximately 30 native and non-native speakers. Every speaker produces around 10 repetitions of each target word. In addition, the corpus provides human evaluation of the pronunciation quality in the range 1-7, 7 being the best. Eight evaluators assessed 4459 acoustic word realizations, from the native and non-native speakers and across all the words in the vocabulary.

3.1. Baseline Experiments

Native and non-native phone and word models were trained using the HTK 3.0 toolkit. We evaluated word recognition in an independent test set for the native and non-native material individually, to analyze the quality of the training process in an independent test set. About 90% and 10% of the data was considered for training and testing, respectively. In both scenarios almost perfect classification results where obtained, which can be expected based on the simplicity of the task.

Before evaluating automatic segmental scores, the human inter-agreement was computed. This is an important reference for evaluating automatic pronunciation scoring, given that there is an intrinsic uncertainty in the labeling in this type of data. In our corpus, we used the 8 evaluators to compute the matrix of human inter-agreement – the correlation coefficients between every pair of evaluators, obtained across all 4459 utterances.

One thing interesting to note is that the level of inter-agreement for the more consistent evaluators is in the range of 0.5-0.67. This range needs to be considered as the upper bound to evaluate automatic assessment techniques. Also these values are relatively low, which can be explained by the fact that the evaluation is at the word level, which makes it particularly difficult. Furthermore, some of the words are mono-syllabic and consequently the evaluators have limited acoustic evidence for doing the assessment. For the evaluation of automatic scores we focus on the three most consistent evaluators.

3.2. Segmental Score Analysis

For implementing automatic pronunciation quality scores, different approaches were considered. The different techniques were based on considering different types of target and background acoustic models. The approaches can be categorized as:

Log-likelihood scores:

log-like-global - log-likelihood score for all the segments normalized at the word level for the number of frames

log-like-local – duration-normalized segmental log-likelihood scores, normalized by the number of segments at the word level

Log-likelihood ratio scores:

phone_nat_non_nat - native phone-level acoustic models for the target and non-native phone-level acoustic models for the background

phone_nat_garb - native phone-level models for the target and phone-level background model trained with all the native acoustic units

word_nat_non_nat - native word-level models for the target and non-native word level models for the background

word_nat_garb - native word-level acoustic models for the target and word-level background models trained with all the native acoustic units

The correlation coefficients between the list of evaluators and the automatic pronunciation quality scores were computed. From these preliminary results we derived some important conclusions. First, log-likelihood ratio scores provide better correlation with human evaluation than correlation obtained with log-likelihood scores alone, which is consistent with previous studies in which it was shown that scores based on confidence measures always provide better performance [3]. On the other hand, word-level scores provide a significant improvement with respect to the use of phone-level acoustic models under the same assessment technique. This last point confirms our well-founded assumption that using longer-context models improves the acoustic description of what the human evaluators consider a good target pronunciation, and consequently improves the performance of the system. The problem is that most real scenarios do not have enough instances of the individual word for training word-level acoustic models. Hence, phone-level models remain the standard. Finally, from these results we can conclude that the use of a proper background model is also a significant aspect of the design of an automatic assessment system. In particular, the results indicated that considering the non-native target models as the background provides some improvement in performance. This improvement is very significant for the case of phone-level models, considerably reducing the performance gap between phone-based and word-based scores. Also of note was that performance obtained for

	Racetrack	Understand	Forgetful	Paper	Wood	Drag	Typical	Thing
native/non divergence	711.76	700.85	548.49	537.59	479.53	467.81	466.99	444.48
phone_nat_non_nat	0.746	0.498	0.607	0.611	0.668	0.666	0.61	0.362
phone_nat_garb	0.715	0.435	0.512	0.525	0.638	0.62	0.473	0.212
word_nat_non_nat	0.656	0.709	0.645	0.65	0.73	0.705	0.672	0.504
word_nat_garb	0.631	0.695	0.624	0.625	0.686	0.698	0.678	0.475

Table 2. Average correlation of automatic and human scores for various words and scoring schemes, in comparison with divergence between native and non-native word-level models.

word-level and some of the phone-level scores are in the range of inter-agreement observed between the consistent evaluators, which is also consistent with results presented in previous studies, showing that segmental log-likelihood ratio scores can approximate the level of self-agreement of human evaluators [3,4].

3.3. Acoustic Distance Measure Analysis

In this section we discuss performance of the different automatic scores as a function of the target word used for the evaluation. The idea is to show experimental evidence that supports the notion that the evaluation vocabulary is a critical design aspect for automatic assessment. In particular, we expect that the performance of automatic scores will improve if we have higher acoustic differences, on average, between native and non-native acoustic production of a target word in the assessment system.

We use the Kullback-Leibler divergence (KLD) as a natural acoustic discrimination measure for comparing native and non-nonnative HMMs associated with the same target acoustic unit. The KLD does not have a closed-form expression for left-to-right HMMs, so a numerical approximation based on Monte Carlo simulation was used to compute its symmetrical extension, the divergence, for every pair of word-level models in our target vocabulary. Table 2 presents the word-dependent average correlation with respect to the set of consistent evaluators. Those results clearly confirm our assumption. For the set of words that have the highest acoustic dissimilarity, better performance is obtained in estimating human judgment of pronunciation quality. This tendency is almost consistent across the different scores evaluated.

4. SYLLABLE STRESS DETECTION

Awareness of proper lexical stress is very important to students of a foreign language. In English, for instance, misplaced syllabic stress can alter a word's part of speech (in the case of "rebel" or "insult") or even change the word's meaning entirely (as with "content" or "contract"). So any interactive computer program for language learners needs to be able to automatically detect a non-canonical stress pattern at least as well as a human tutor would.

The data we used for these experiments came once again from the ISLE Corpus compiled at the University of Leeds [1], as in Section 2. Utterance prompts were complete sentences written by design to highlight certain difficulties English learners typically encounter, both in phonemic pronunciation and in recognizing variations in primary lexical stress (e.g. "project" when used as a noun vs. when used as a verb). The recordings were automatically tagged for canonical forms by a forced-aligner, then corrected to reflect the speaker's pronunciation by a team of five linguists, who also added labels for each word's syllable of primary acoustic stress (compared to the canonical).

4.1. Prosodic Features

The three basic prosodic features chosen for baseline experiments were mean values of f0 and energy over the syllable nucleus (normalized by the respective mean values over the entire word) and the nucleus duration (normalized by the mean nucleus duration over all syllables in that word). Normalizing in this way preserves the word context information and is in keeping with the fundamental idea that we're not so much interested in classifying each syllable separately, but rather we intend to compare characteristics of all syllables in a word and choose one as the location of primary stress.

We also included several features related to the f0 slope, which proved to be closely correlated with syllabic stress. The importance of these slope-derived features is in their potential to capture higher-level pitch information, to model the rapid rate of f0 and energy changes that correlate with stress but are in some sense independent of the mean f0 value. These features are inspired by the ones used for pronunciation evaluation and speaker recognition in [5] and [6], respectively. For reasons similar to those of the slope-derived features, we also included two features derived from the syllable's pitch and energy range. All features are discussed in detail in [7].

Inclusion of these slope- and range-related features serves to make our model of syllabic "stress" a combination of [8]'s close distinctions between stress and pitch-accent, which we argue is necessary for language learning purposes. The syllable nuclei durations were derived from automatic results of forced alignments based on transcriptions of each recording's utterance prompt. So it was possible for us to incorporate higher-level contextual information to optimize normalization of our syllable duration feature, which was necessary especially because we used the syllable nucleus in place of the syllable itself. [9] presents a list of linguistic rules with which to further normalize vowel durations based on that vowel's word- and phrase-level context. These rules seem to be derived from empirically calculated average durational trends in pronunciation.

4.2. Experiments and Results

We began the classification process by considering this a two-class problem: we started by classifying each syllable individually as stressed or unstressed, without regard for within-word information.

For training data, we used 13 native Italian speakers (a total of 7086 syllable nuclei, taken only from polysyllabic words), and 12 native German speakers (7878 syllable nuclei instances), all taken from the ISLE corpus described above in Section 2. Classifiers for

	Italian			German		
	syllable	w/ word info	word	syllable	w/ word info	word
3 basic features	75.63	76.51	83.39	80.26	82.00	87.80
ranges	56.74	57.34	64.96	61.96	62.46	69.69
slopes	62.82	66.91	73.10	67.43	69.15	75.60
basic + ranges	78.16	78.31	83.87	82.07	82.25	87.41
basic + slopes	82.48	84.01	87.61	85.49	85.73	89.14
all 10 features	82.57	83.17	86.75	85.57	85.81	88.81

Table 3. Syllable stress detection accuracy for individual syllables with and without word information, and inter-word accuracy based on the percentage of complete words in which all syllables were post-classified correctly.

the Italian and German students were trained and tested separately, since we may assume the classifier has prior knowledge of the registered student's native language. The test set was comprised of the remaining 10 Italian and 11 German speakers. The classifier chosen was a quadratic Bayes discriminant function, assuming Normal distributions.

After individually classifying every syllable nucleus as stressed or unstressed, we sought to improve accuracy by including information about intra-word stress results. By definition, no word can have more or less than one syllable of primary stress, and the ISLE corpus is labeled accordingly. So in the words for which our classifier assigned more than one primary stress, we kept the one with the best posterior probability returned by the classifier, and post-classified the other ones as unstressed. And in words with no stressed syllable results, we chose the one unstressed syllable with the worst posterior probability and post-classified it as stressed.

Results using different feature sets are shown in Table 3. All results listed include [9]'s contextual rules for normalizing vowel durations. From Table 3 we can see that the classifier performed slightly better overall for the Germans than for the Italians. This might be due to the fact that the contextual rules for normalizing durations in [9] were taken from a study in American English, and German is linguistically more closely related to English than Italian is. However, testing the German speakers on the Italian-trained models (and vice-versa) did not result in a significant decline in accuracy. This seems to indicate that, with more diverse training data, one should be able to generalize these models for all English learners, regardless of their native language (or at least define models in broader linguistic groups).

The human experts did not tag each syllable individually, without regard for other syllables in the same word. No, they listened to each word separately and picked one syllable as the location of primary stress. So in the end, the best measure of our method's performance is really the word accuracy ratings – the percentage of words in which all syllables were classified correctly. Now inter-human agreement in linguistic labeling is commonly held (e.g. by [5,8]) to be about 80%. So, by Table 3, even a few of our sub-optimal feature selections performed as well as a human labeler would. And our results using baseline features were comparable with that of similar features employed in [8].

As far as the features go, the baseline experiments using only mean f0, mean energy, and duration already yielded a word accuracy rate of better than 80%. Adding features related to the f0 slope and the f0 and energy ranges was necessary only to push the individual syllable accuracy above 80%, in keeping with the rating convention of [8]. Though the range-related features did add some improvement over the baseline, the classifier performed better when only the baseline and slope-related features were used.

As a supplementary experiment, we calculated the classifier output's accuracy when compared with the canonical stress pattern (not the human-tagged stress labels), by way of generating some kind of crude overall pronunciation score for each speaker (a true score would include more than just accuracy of stress placement). We also calculated comparable objective "human" scores based on the agreement between the hand-tagged stress data and the canonical stress marks. The correlation between the automatic scores and the human scores for the 10 speakers in the Italian test set was 0.7998 for syllables, 0.8087 for words, which beats the inter-human correlation in assigning general pronunciation scores to speakers in this corpus (it was no better than 0.7, as reported in [1]). This speaks well for our method's applicability to pronunciation evaluation on the whole.

5. TWO EXAMPLE APPLICATIONS

5.1. Tball Literacy Assessment

Automatically assessing a child's literacy skills is a complex problem. Based on a read-aloud speech signal along with prior knowledge of the expected target utterance, we can infer quite a bit – the child's confidence in reading, his level of fluency or comfort, even the influence and degree of non-native phonetics – just as a real reading tutor could. But how we can distinguish a mispronunciation based on underdeveloped reading skills from one caused by a non-native accent or speaker-dependent speech production difficulties is another matter.

In an ASR task such as this, a given mispronunciation cannot be presumed to be the sole result of any one source. The child who, when prompted with the word /f ay n d/ ("find"), reads aloud something that sounds like /f ih n d/, probably does so because of an unfamiliarity with the orthographic conventions of English letter-to-sound rules. However, first-graders of Mexican-American background (as populate much of the Los Angeles public school system) might read the target word "two" more like "do" because of the shorter Voice Onset Time in Spanish-accented stops, but chances are this pronunciation variant would not be the product of poor reading skills, and therefore should not be assessed as such. Combine this ambiguity with an easily confusable wordlist (typical Grade 1 words: well/will, saw/so, etc.) and the high age-dependent variability of children's speech, and you have an evaluation problem for which simple word-level recognition grammars and traditional log-likelihood ratio thresholding will not suffice.

Additionally, experts in child literacy don't always agree on what constitutes an acceptable mispronunciation by a speaker with a nonnative accent, so there is always a degree of uncertainty in these assigned class labels.

The data used in this study comes from the Tball Corpus [10], which was gathered at Los Angeles area schools and motivated by a long-term goal to develop automatic literacy assessment software for elementary school teachers. The particular subset we used is the Grade 1 word list recordings, consisting of 2076 one-word utterances from roughly an equal number of boys and girls, ages 5-8, over a 51-word vocabulary typical of first grade reading ability.

5.1.1. Pronunciation Verification

Section 3 describes how word-level verification is often performed, though it has been shown to be less than useful in all but the simplest of recognition tasks, since the threshold is not easily generalized for a large vocabulary; depending on their phonetic properties, certain words will require a verification threshold farther from the filler model than others. One suggested improvement [11] is to use a unique filler model for each target word, one that omits any instances of the target word in question during the training stage, though this necessitates retraining acoustic models each time the reading list vocabulary is changed.

Clearly this classification task demands more features, and perhaps a more complex classification algorithm. Sources such as [12] suggest deriving new acoustic scores based on a recognition grammar over the entire task vocabulary, rather than from fixed alignment of the target word and global filler model. Likelihood scores for linguistically close pronunciations will serve as a discriminative foil for the target pronunciation's acoustic model, and the distant words need not be considered, and will not be recognized except in the case of an extreme mispronunciation. In this way we can then estimate the filler model dynamically without severe training overhead, and focus on improving performance in the case more commonly seen in pronunciation verification – that of false acceptance.

As for the classifier, a linear threshold is a good place to start, but the framework needs to be augmented to account for cases where, for example, the log-likelihood ratio is relatively high but the target likelihood component is not, or the target word is not recognized despite the high score computed upon alignment with the target model (both cases and many more idiosyncratic were well-represented in the data). For this reason, and because of its acceptance of both binary and continuous features and its easy interpretability, we decided to use a decision tree classifier for our pronunciation assessment.

We also propose using human evaluation knowledge besides on the transcription level. This is more or less an unsupervised learning task – we do not know in advance the "true" class labels, acceptable or unacceptable, for any of the pronunciations, because even expert labelers cannot always agree when evaluating pronunciation. So to train an accurate decision tree, we'll need to somehow estimate the true class labels using human evaluations – the same human evaluations collected for purposes of comparison with our automatic classification results. Our method intends to demonstrate an improvement in the classification results when said classifier is informed by human evaluations.

Our evaluation consisted of a set of 102 utterances, two examples of each target word from the Grade 1 word list, representative of typical canonical and noncanonical pronunciations, respectively. Our 20 evaluators – 3 of them teachers, 8 of them native American English speakers, all of varying degrees of Spanish language fluency – were asked to mark each item as "acceptable" or "unacceptable" based on the child's pronunciation of the target word.

Though the expert teacher agreement is numerically higher than the non-teacher class, we found with 95% confidence that the teachers did not have statistically higher inter-agreement than the non-teachers. The same was true of the native vs. non-native agreement means. This indicates that experts and native speakers do not necessarily perceive pronunciation with dramatically more agreement than anyone else.

5.1.2. Feature Selection

For baseline experiments with the traditional approach, we used only the traditional confidence measure in Section 3, calculated based on the likelihood score after alignment with the target word models normalized by two different filler models (for comparison): the general word-level filler, and a dynamic filler derived from the likelihood score of the recognized word. For the baseline threshold classifiers, these scores were obtained with the dictionary of canonical pronunciations – no prior linguistic rules were used.

Training of the decision tree classifier included all these baseline features – target word likelihood score, recognized word likelihood score, word-level filler, dynamic filler, and all combinations of likelihood ratios – though a dictionary of pronunciation variants was used to obtain target word alignment scores, and the dynamic filler was calculated by averaging the likelihood scores of the 20-best results. The best recognition result was included as a binary feature (1 if it matched the target word, 0 if it didn't), and the percentage of the 20-best results which matched the target was also included as a feature. The differences between comparable target and recognized word likelihood ratios were included as well. And the student's response time and word duration were also used as features, as they may be indicative of pronunciation fluency.

5.1.3. Experimental Setup

The threshold imposed on the baseline log-likelihood ratios was determined empirically over a wide range of possible thresholds, the best one being selected based on highest agreement with the human evaluations. We explored two techniques for assigning class labels to the decision tree's training set: one, take a majority "vote" of the human evaluations for what the true class should be (these voted class labels had 93% average agreement with the three expert teacher evaluators); two, use the word-level transcriptions so that if the transcribed word matches the target word, we put it in the acceptance class, otherwise it's in the rejection class. The latter method does not necessarily agree in the pronunciation verification case, since an acceptable mispronunciation might generate a different dictionary word as surface form (as with do/two); however, the transcription class labels were found to agree with the voted ones 95% of the time, so the method seemed a valid choice – we can think of the transcriptions as data from another expert human evaluator (and they are, in fact), with which to compare our automatic results. These decision tree training methods were compared using a leave-one-out crossvalidation procedure over the entire evaluation set.

5.1.4. Results and Discussion

The results of our four classifiers are enumerated in Table 4,

		Kappa	*P*(*agreement*)
inter-evaluator	*teachers*	0.77	0.89
	voted	0.85	0.93
	all	0.69	0.85
threshold : general (baseline)	*teachers*	0.59	0.80
	voted	0.66	0.83
	all	0.55	0.78
threshold : dynamic	*teachers*	0.72	0.87
	voted	0.82	0.91
	all	0.67	0.84
tree : voting	*teachers*	0.72	0.86
	voted	0.80	0.90
	all	0.68	0.84
tree : transcripts	*teachers*	**0.72**	**0.86**
	voted	**0.82**	**0.91**
	all	**0.67**	**0.84**

Table 4. Mean Kappa and agreement statistics for human evaluators and four classifiers, compared with expert teacher evaluators, the voted class labels, and averaged over all evaluators.

alongside statistics for inter-evaluator agreement. Each algorithm is compared to the expert teacher evaluators, the voted approximation of the "true" class labels, and an average with respect to all 20 evaluators. In the context of this work, the log-likelihood ratio threshold classifier with the general filler model (*threshold : general*) can be considered as a baseline for automatic verification. And, as expected, it performed the poorest of the four algorithms. The other three all had very similar performance, and came well within the 6% standard deviation of the inter-evaluator agreement scores. The voted class labels (an approximation of what the "true" classes may be) agree with all human evaluators 93% of the time, on average, so a 90-91% agreement with the voted class labels indicates that these classification algorithms perform about as well as a human evaluator. And since the teachers agree among themselves 89% of the time, our 86-87% agreement with the teachers suggests these automatic methods can serve about as well as an expert evaluator. In outperforming the baseline, the other classifiers demonstrated that expert prior knowledge, in the form of human evaluations and acceptable pronunciation variants, can dramatically improve classifier performance.

Of the three best classifiers, the simple threshold classifier with dynamic filler model (*threshold : dynamic*) performed almost as well as the more complex decision trees. However, to set the optimal decision threshold for both of the traditional classification schemes, we explored a range of thresholds and chose the one with the highest agreement with human evaluations. Consequently, we can say that we have an over-optimistic setting for the traditional threshold systems, because we used test set performance information to iteratively perfect the classification of the test set itself. Whereas the decision tree results are based on a leave-one-out crossvalidation procedure which keeps the training and test instances separate and relies on human evaluations only as training set labels. But the high *threshold : dynamic* results suggest that the large number of extraneous features may not be necessary. After pruning, the decision trees were found to branch on only three of the fourteen attributes.

Using the transcript-based class labels to train the decision tree (*tree : transcripts*) resulted in a slightly better average agreement with the human evaluations. This seems to indicate that the human evaluators were choosing to reject a pronunciation if the variant resulted in a new dictionary word, and would accept only what they perceived to be a surface form variant of the target word that did not become an entirely different word, much like word-level transcribers. We can conclude, then, that to provide class labels for our decision tree's training set, we probably only need one expert evaluator: a transcriber.

5.2. Tactical Language Training System

The Tactical Language Training System uses an Automated Speech Recognition (ASR)-driven pedagogical environment to guide the learner in rapid acquisition of basic language and cultural skills [13]. We currently have systems built for Levantine and Iraqi Arabic, and have a Pashto system under development. Language learners perform a variety of interactive exercises, speaking to the computer and getting feedback on their choice of responses and pronunciation. They also practice in an interactive game in which they must communicate with non-player characters, giving learners a task-based method of learning the language.

5.2.1. Analysis of Learner Speech Errors

The TLTS incorporates two speech-enabled learning environments: an interactive game called the Mission Practice Environment (MPE) that simulates conversations with native speakers, and an intelligent tutoring system called the Mission Skill Builder (MSB) for acquiring and practicing communicative skills. The speech error analysis module is responsible for providing users with feedback on their performance in both these environments.

The objective of the error analysis module is to categorize the ASR-processed learner speech signal. The first task of this module is to model the phonological errors using offline training data. Instead of a rule-based system, the TLTS uses a Naive Bayesian classifier that was trained on native speaker annotation of learner speech. To train the phonological error model, we used a corpus of 1893 non-native speaker pronunciations, gathered from 7 male non-native speakers with no prior Levantine experience saying 188 distinct words. From this corpus, we apply techniques similar to those used in machine translation to find "translations" from phonemes in native speech to phonemes in learner speech. More specifically, we use co-occurrence statistics to populate a Bayesian model of phoneme-to-phoneme (and phoneme-to-nil) n-gram mappings, and use these to generate a noisy-channel system modeling learner mistakes. In the next stage of error analysis we use a number of factors such as ASR confidence in error detection, derived confidence as inferred from past learner history, and intrinsic characteristics of the error committed are taken into account to contextualize learner errors and re-rank them.

First, raw ASR confidence gives us an idea of error severity. We boost raw ASR confidence by considering results in the context of the learner's performance history. Positive evidence of the learner having made a specific mistake in the past boosts our confidence of that error in current detection, and likewise trends of performing a problematic speech unit correctly lowers our confidence. Subsequently the characteristics of the errors committed, as taken in cultural and listener context are judged for

their severity. Finally, errors are considered extremely severe when collision with other words causes the learner to break social taboos (e.g. mixing up the very phonetically similar words /raaxid/ ["terrible"] and /raa'id/ ["major"]).

5.2.2. ASR for Mission Practice Environment (MPE)

The speech recognition system was implemented using the Cambridge HTK toolkit. The system was trained on a modern standard Arabic dataset with around 10 hours of native speech. A mapping from modern standard Arabic phone set to Levantine Arabic phone set was used for this purpose.

The ASR training and decoding process differs for the two TLTS learning environments in tune with their distinct objectives. The goal of the ASR in MPE is to enable the user to communicate with virtual agents using speech. Thus, the primary objective for this task is to do robust recognition of learner speech to enable free form, naturalistic interaction. The MPE grammar thus includes utterances and their common morpho-syntactic variants in a finite state grammar. The MPE acoustic models are built from native speech and un-annotated non-native speech.

5.2.3. ASR for Mission Skill Builder (MSB)

The MSB grammar is tuned for detecting pronunciation variants and was generated using the error analysis grammar and generation method described above. The MSB acoustic models are built from native speech and annotated non-native speech.

Error disambiguation requires a measure of ASR confidence and hypothesis verification for error identification. This is carried out using a two-step thresholding procedure. In the first step we decode the utterance with a grammar containing the correct canonical pronunciation of all the utterances in the system and a reference single path grammar containing just the intended utterance. If the output of the recognizer is the same as the intended utterance we proceed to the next step. In case the two differ we compare the acoustic score of the utterance grammar with the reference. If the difference between the two scores is higher than a leniency threshold, the utterance is judged to be out of the intended vocabulary of the learner, and the system thus generates a generic incorrect utterance feedback to the user. If this is not the case, the system diagnoses specific learner errors. For this diagnosis we decode the utterance with its corresponding phone confusion grammar whose generation is described above. If the recognizer selects the canonical output from the error analysis grammar or the difference in acoustic score between the canonical and the best path is less than a threshold, we provide the user with a generic positive feedback. Otherwise feedback is generated using the error disambiguation procedure described above. ASR confidence for disambiguation is taken to be the difference in system confidence between the recognized variant, the output from the utterance recognizer, and the canonical utterance.

The thresholds were determined using 500 annotated utterances as a heldout set. An analysis of the recognizer errors indicates that words with lengthened vowel English transliteration (ii, aa) needed significantly higher leniency threshold than other words. The thresholds are adjusted in the system based on user expertise/difficulty level and the pedagogical history.

6. CONCLUSION

These methods of locating an expected pronunciation mistake represent several preliminary steps in the complex pronunciation evaluation task. Once the presence of a recurring error has been identified, a robust language learning system should provide physically meaningful instruction to the student as to a proper articulatory reconfiguration ("keep your lips rounded," for instance). Future work in this area could focus on optimizing the usefulness of this feedback.

7. ACKNOWLEDGMENTS

This work is supported by grants from the National Science Foundation, DARPA, and ONR.

8. REFERENCES

[1] E. Atwell, P. Howarth, C. Souter, "The ISLE Corpus: Italian and German Spoken Learner's English," *ICAME Journal*, Vol. 27, pp. 5-18, 2003.

[2] M. Richardson, J. Bilmes, and C. Diorio, "Hidden-Articulator Markov Models for Speech Recognition," *ASR2000*.

[3] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. "Combination of Machine Scores for Automatic Grading of Pronunciation Quality." *Speech Communication*, 30(2–3):121–130, Feb 2000.

[4] Cucchiarini, C., Strik, H. and Boves, L., "Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms." *Speech Communication, 30*, pp. 109-119, 2000.

[5] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, K. Sonmez, "Evaluation of Speaker's Degree of Nativeness Using Text-Independent Prosodic Features," *Proc. Of the Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[6] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, B. Xiang, "Using Prosodic and Conversational Features for High-Performance Speaker Recognition," *Johns Hopkins University Workshop 2002*.

[7] J. Tepperman and S. Narayanan, "Automatic Syllable Stress Detection for Pronunciation Evaluation of Language Learners," *Proc. ICASSP'05*, Philadelphia, 2005.

[8] F. Tamburini, "Prosodic Prominence Detection in Speech," *Proc. ISSPA2003*, Paris, pp. 385-388.

[9] R. Delmonte, M. Petrea, and C. Bacalu, "SLIM Prosodic Module for Learning Activities in a Foreign Language," *Proc. ESCA*, Eurospeech97, Rhodes, Vol. 2, pp. 669-672.

[10] 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. Eurospeech*, Lisbon, 2005.

[11] P. Ramesh, C.-H. Lee, and B.-H. Juang, "Context Dependent Anti Subword Modeling for Utterance Verification," in *Proc. of ICSLP*, Sydney, 1998.

[12] J. Caminero, C. de la Torre, L. Villarrubia, C. Matin, and L. Hernandez, "On-line garbage modeling with discriminant analysis for utterance verification," in *Proc. of ICSLP*, Philadelphia, 1996.

[13] L. Johnson, S. Marsella, N. Mote, M. Si, H. Vilhjalmsson, S. Wu. "Balanced Perception and Action in the Tactical Language Training System", In InSTIL, 2004.