Towards Automatic Scoring of Non-Native Spontaneous Speech

Klaus Zechner and **Isaac I. Bejar**
Educational Testing Service
Princeton, NJ, USA
(kzechner,ibejar)@ets.org

Abstract

This paper investigates the feasibility of automated scoring of spoken English proficiency of non-native speakers. Unlike existing automated assessments of spoken English, our data consists of spontaneous spoken responses to complex test items. We perform both a quantitative and a qualitative analysis of these features using two different machine learning approaches. (1) We use support vector machines to produce a score and evaluate it with respect to a mode baseline and to human rater agreement. We find that scoring based on support vector machines yields accuracies approaching inter-rater agreement in some cases. (2) We use classification and regression trees to understand the role of different features and feature classes in the characterization of speaking proficiency by human scorers. Our analysis shows that across all the test items most or all the feature classes are used in the nodes of the trees suggesting that the scores are, appropriately, a combination of multiple components of speaking proficiency. Future research will concentrate on extending the set of features and introducing new feature classes to arrive at a scoring model that comprises additional relevant aspects of speaking proficiency.

1 Introduction

While automated scoring of open-ended written discourse has been approached by several groups recently (Rudner & Gagne, 2001; Shermis & Burstein, 2003), automated scoring of spontaneous spoken language has proven to be more challenging and complex. Spoken language tests are still mostly scored by human raters. However, several systems exist that score different aspects of spoken language; (Bernstein, 1999; C. Cucchiarini, H. Strik, & L. Boves, 1997a; Franco et al., 2000). Our work departs from previous research in that our goal is to study the feasibility of automating scoring for *spontaneous speech*, that is, when the spoken text is not known in advance.

We approach scoring here as the characterization of a speaker's oral proficiency based on features that can be extracted from a spoken response to a well defined test question by means of automatic speech recognition (ASR). We further approach scoring as the construction of a mapping from a set of features to a score scale, in our case five discrete scores from 1 (least proficient) to 5 (most proficient). The set of features and the specific mapping are motivated by the concept of communicative competence (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972). This means that the features in the scoring system we are developing are meant to characterize specific components of communicative competence, such as mastery of pronunciation, fluency, prosodic, lexical, grammatical and pragmatical subskills. The selection of features is guided by an understanding of the nature of speaking proficiency. We rely on the scoring behavior of judges to evaluate the features (section 8) as well as a convenient criterion for evaluating the feasibility of automated scoring based on those features (section 7). That is, the role of human scorers in this context is to provide a standard for system evaluations (see section 7), as well as to validate specific features and feature classes chosen by the authors (section 8). We use support vector machines (SVMs)

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 216–223,
New York, June 2006. ©2006 Association for Computational Linguistics

to determine how well the features recover human scores. We collect performance data under three different conditions, where features are either based on actual recognizer output or on forced alignment. (Forced alignment describes a procedure in speech recognition where the recognizer is looking for the most likely path through the Hidden Markov Models given a transcription of the speech file by an experienced transcriber. This helps, e.g., in finding start and end times of words or phonemes.) We then use classification and regression trees (CART) as a means to evaluate the relative importance and salience of our features. When the classification criterion is a human score, as is the case in this study, an inspection of the CART tree can give us insights into the feature preferences a human judge might have in deciding on a score.

The organization of this paper is as follows: first, we discuss related work in spoken language scoring. Next, we introduce the data of our study and the speech recognizer used. In section 5 we describe features we used for this study. Section 6 describes the agreement among raters for this data. Section 7 describes the SVM analysis, section 8 the CART analysis. This is followed by a discussion and then finally by conclusions and an outlook on future work.

2 Related work

There has been previous work to characterize aspects of communicative competence such as fluency, pronunciation, and prosody. (Franco et al., 2000) present a system for automatic evaluation of pronunciation performance on a phone level and a sentence level of native and non-native speakers of English and other languages (EduSpeak). Candidates read English text and a forced alignment between the speech signal and the ideal path through the Hidden Markov Model (HMM) was computed. Next, the log posterior probabilities for pronouncing a certain phone at a certain position in the signal were computed to achieve a local pronunciation score. These scores are then combined with other automatically derived measures such as the rate of speech (number of words per second) or the duration of phonemes to yield global scores.

(C. Cucchiarini, S. Strik, & L. Boves, 1997b)) and (Cucchiarini et al., 1997a)) describe a system for Dutch pronunciation scoring along similar lines. Their feature set, however, is more extensive and contains, in addition to log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. In an evaluation, they find good agreement between human scores and machine scores.

(Bernstein, 1999)) presents a test for spoken English (SET-10) that has the following types of items: reading, repetition, fill-in-the-blank, opposites and open-ended answers. All types except for the last are scored automatically and a score is reported that can be interpreted as an indicator of how native-like a speaker's speech is. In (Bernstein, DeJong, Pisoni, & Townshend, 2000), an experiment is performed to establish the generalizability of the SET-10 test. It is shown that this test's output can successfully be mapped to the Council of Europe's Framework for describing second language proficiency (North, 2000). This paper further reports on studies done to correlate the SET-10 with two other tests of English proficiency, which are scored by humans and where communicative competence is tested for. Correlations were found to be between 0.73 and 0.88.

3 Data

The data we are using for the experiments in this paper comes from a 2002 trial administration of TOEFLiBT® (Test Of English as a Foreign Language—internet-Based Test) for non-native speakers (LanguEdge ™). Item responses were transcribed from the digital recording of each response. In all there are 927 responses from 171 speakers. Of these, 798 recordings were from one of five main test items, identified as P-A, P-C, P-T, P-E and P-W. The remaining 129 responses were from other questions. As reported below, we use all 927 responses in the adaptation of the speech recognizer but the SVM and CART analyses are based on the 798 responses to the five test items. Of the five test items, three are *independent* tasks (P-A, P-C, P-T) where candidates have to talk freely about a certain topic for 60 seconds. An example might be "Tell me about your favorite teacher." Two of the test items are *integrated* tasks (P-E, P-W) where candidates first read or listen to some material to which they then have to relate in their responses (90 seconds speaking time). An example might be that the candidates listen to a conversational argument about studying at home vs. studying abroad and then are asked to summarize the advantages and disadvantages of both points of view.

The textual transcription of our data set contains about 123,000 words and the audio files are in WAV format and recorded with a sampling rate of 11025Hz and a resolution of 8 bit.

For the purpose of adaptation of the speech recognizer, we split the full data (927 recordings) into a training (596) and a test set (331 recordings). For the CART and SVM analyses we have 511 files in the *train* and 287 files in the *eval* set, summing up to 798. (Both data sets are subsets from the ASR adaptation training and test sets, respectively.) The transcriptions of the audio files were done according to a transcription manual derived from the German VerbMobil project (Burger, 1995). A wide variety of disfluencies are accounted for, such as, e.g., false starts, repetitions, fillers, or incomplete words. One single annotator transcribed the complete corpus; for the purpose of testing inter-coder agreement, a second annotator transcribed about 100 audio files, which were randomly selected from the complete set of 927 files. The disagreement between annotators, measured as word error rate (WER = (substitutions + deletions + insertions) / (substitutions + deletions + correct)) was slightly above 20% (only lexical entries were measured here). This is markedly more disagreement than in other corpora, e.g., in SwitchBoard (Meteer & al., 1995) where disagreements in the order of 5% are reported, but we have non-native speech from speakers at different levels of proficiency which is more challenging to transcribe.

4 Speech recognition system

Our speech recognizer is a gender-independent Hidden Markov Model system that was trained on 200 hours of dictation data by native speakers of English. 32 cepstral coefficients are used; the dictionary has about 30,000 entries. The sampling rate of the recognizer is 16000Hz as opposed to 11025Hz for the LanguEdge™ corpus. The recognizer can accommodate this difference internally by up-sampling the input data stream.

As our speech recognition system was trained on data quite different from our application (dictation vs. spontaneous speech and native vs. non-native speakers) we adapted the system to the LanguEdge ™ corpus. We were able to increase word accuracy on the unseen test set from 15% before adaptation to 33% in the fully adapted model (both acoustic and language model adaptation).

5 Features

Our feature set, partly inspired by (Cucchiarini et al., 1997a), focuses on low-level fluency features, but also includes some features related to lexical sophistication and to content. The feature set also stems, in part, from the written guidelines used by human raters for scoring this data. The features can be categorized as follows: (1) Length measures, (2) lexical sophistication measures, (3) fluency measures, (4) rate measures, and (5) content measures. Table 1 renders a complete list of the features we computed, along with a brief explanation. We do not claim these features to provide a full characterization of communicative competence; they should be seen as a first step in this direction. The goal of the research is to gradually build such a set of features to eventually achieve as large a coverage of communicative competence as possible. The features are computed based on the output of the recognition engine based on either forced alignment or on actual recognition. The output consists of (a) start and end time of every token and hence potential silence in between (used for most features); (b) identity of filler words (for disfluency-related features); and (c) word identity (for content features).

Lexical counts and length measures	
Segdur	Total duration in seconds of all the utterances
Numutt	Number of utterances in the response
Numwds	Total number of word forms in the speech sample
Numdff	Number of disfluencies (fillers)
Numtok	Number of tokens = Numwds+Numdff
Lexical sophistication	
Types	Number of unique word forms in the speech sample
Ttratio	Ratio Types/Numtok (type-token ratio, TTR)
Fluency measures **(based on pause information)**	
Numsil	Number of silences, excluding silences between utterances
Silpwd	Ratio Numsil/Numwds
Silmean	Mean duration in seconds of all silences in a response to a test item
Silstddv	Standard deviation of silence duration
Rate measures	
Wpsec	Number of words per second
Dpsec.	Number of disfluencies per second
Tpsec	Number of types per second
Silpsec.	Number of silences per second
Content measures	We first compute test-item-specific word vectors with the frequency counts of all words occurring in the *train* set for each test item (wvec_testitem). Then we generate for every item response a word vector in kind (wvec_response) and finally compute the inner product to yield a similarity score: sim = wvec_testitem*wvec_response
Cvfull	wvec_testitem*wvec_response
6 other Cv*-features	As Cvfull but measure similarity to a subset of wvec_testitem, based on the scores in the *train* set (e.g., “all responses with score 1”)
Cvlennorm	Length-normalized Cvfull: Cvfull/Numwds

Table 1: List of features with definitions.

6 Inter-rater agreement

The training and scoring procedures followed standard practices in large scale testing. Scorers are trained to apply the scoring standards that have been previously agreed upon by the developers of the test. The training takes the form of discussing multiple instances of responses at each score level. The scoring of the responses used for training other raters is done by more experienced scorers working closely with the designers of the test.

All the 927 speaking samples (see section 3) were rated once by one of several expert raters, which we call Rater1. A second rating was obtained for approximately one half (454) of the speaking samples, which we call Rater2. We computed the exact agreement for all Rater1-Rater2 pairs for all five test items and report the results in the last column of Table 2. Overall, the exact agreement was about 49% and the kappa coefficient 0.34. These are rather low numbers and certainly demonstrate the difficulty of the rating task for humans. Inter-rater agreement for integrated tasks is lower than for independent tasks. We conjecture that this is related to the dual nature of scoring integrated tasks: for one, the communicative competence per se needs to be assessed, but on the other hand so does the correct interpretation of the written or auditory stimulus material. The low agreement in general is also understandable since the number of feature dimensions that have to be mentally integrated pose a significant cognitive load for judges.[1]

7 SVM models

As we have mentioned earlier, the rationale behind using support vector machines for score prediction is to yield a quantitative analysis of how well our features would work in an actual scoring system, measured against human expert raters. The choice of the particular classifier being SVMs was due to their superior performance in many machine learning tasks.

7.1 Support vector machines

Support vector machines (SVMs) were introduced by (Vapnik, 1995) as an instantiation of his approach to model regularization. They attempt to solve a multivariate discrete classification problem where an n-dimensional hyperplane separates the input vectors into, in the simplest case, two distinct classes. The optimal hyperplane is selected to minimize the classification error on the training data, while maintaining a maximally large margin (the distance of any point from the separating hyperplane).

7.2 Experiments

We built five SVM models based on the *train* data, one for each of the five test items. Each model has two versions: (a) based on forced alignment with the true reference, representing the case with 100% word accuracy (align), and (b) based on the actual recognition output hypotheses (hypo). The SVM models were tested on the *eval* data set and there were three test conditions: (1) both training and test conditions derived from forced alignment (align-align); (2) models trained on forced alignment and evaluated based on actual recognition hypotheses (align-hypo; this represented the realistic situation that while human transcriptions are made for the training set, they would turn out to be too costly when the system is running continuously); and (3) both training and evaluation are based on ASR output in recognition mode (hypo-hypo).

We identified the best models by running a set of SVMs with varying cost factors, ranging from 0.01 to 15, and three different kernels: radial basis function, and polynomial, of second degree and of third degree. We selected the best performing models measured on the *train* set and report results with these models on the *eval* set. The cost factor for all three configurations varied between 5 and 15 among the five test items, and as best kernel we found the radial basis function in almost all cases, except for some polynomial kernels in the hypo-hypo configuration

	Mode (% of eval set)	Train : align Eval : align	Train : align Eval : hypo	Train : hypo Eval : hypo	Human Rater Agreement (% of all pairs)
P-A (ind)	34	40.7	33.9	35.9	53
P-C (ind)	53	50.0	55.0	56.7	57
P-T (ind)	38	43.4	18.9	37.7	54
P-E (int)	25	42.1	26.3	47.4	43
P-W (int)	29	34.5	20.7	39.7	42

5

Table 2: Speech scoring: Mode baseline, SVM performance on forced alignment and standard recognition data, and human agreement for all five test items (ind=independent task; int=integrated task).

[1] Inter-human agreement rates for written language, such as essays, are significantly higher, around 70-80% with a 5-point scale (Y.Attali, personal communication). More recently we observed agreement rates of about 60% for spoken test items, but here a 4-point scale was used.

7.3 Results

Table 2 shows the results for the SVM analysis as well as a baseline measure of agreement and the inter rater agreement. The baseline refers to the expected level of agreement with Rater1 by simply assigning the mode of the distribution of scores for a given question, i.e., to always assign the most frequently occurring score on the *train* set. Table 2 also reports the agreement between trained raters. As can be seen the human agreement is consistently higher than the mode agreement but the difference is less for the integrated questions suggesting that humans scorers found those questions more challenging to score consistently.

The other 3 columns of Table 2 report the results for the perfect agreement between a score assigned by the SVM developed for that test question and Rater1 on the *eval* corpus, which was not used in the development of the SVM. We observe that for the align-align configuration, accuracies are all clearly better than the mode baseline, except for P-C, which has an unusually skewed score distribution and therefore a rather high mode baseline. In the align-hypo case, where SVM models were built based on features derived from ASR forced alignment and where these models were tested using ASR output in recognition mode, we see a general drop in performance – again except for P-C – which is to be expected as the training and test data were derived in different ways. Finally, in the hypo-hypo configuration, using ASR recognition output for both training and testing, SVM models are, in comparison to the align-align models, improved for the two integrated tasks but not for the independent tasks, again except for P-C. The SVM classification accuracies for the integrated tasks are in the range of human scorer agreement, which indicates that a performance ceiling may have been reached already. These results suggest that the recovery of scores is more feasible for integrated rather than independent tasks. However, it is also the case that human scorers had more difficulty with the integrated tasks, as discussed in the previous section.

The fact that the classification performance of the hypo-hypo models is not greatly lower than that of the align-align models, and in some cases even higher ---and that with the relatively low word accuracy of 33%---, leads to our conjecture that this could be due to the majority of features being based on measures which do not require a correct word identity such as measures of rate or pauses.

In a recent study (Xi, Zechner, & Bejar, 2006) with a similar speech corpus we found that while the hypo-hypo models are better than the align-align models when using features related to fluency, the converse is true when using word-based vocabulary features.

8 CART models

8.1 Classification and regression trees

Classification and regression trees (CART trees) were introduced by (Breiman, Friedman, Olshen, & Stone, 1984). The goal of a classification tree is to classify the data such that the data in the terminal or classification nodes is as pure as possible meaning all the cases have the same true classification, in the present case a score provided by a human rater, the variable Rater1 above. At the top of the tree all the data is available and is split into two groups based on a split of one of the features available. Each split is treated in the same manner until no further splits are possible, in which case a terminal node has been reached.

8.2 Tree analysis

For each of the five test items described above we estimated a classification tree using as independent variables the features described in Table 1 and as the dependent variable a human score. The trees were built on the *train* set. Table 3 shows the distribution of features in the CART tree nodes of the five test items (rows) based on feature classes (columns). For P-A, for example, it can be seen that three of the feature classes have a count greater than 0. The last column shows the number of classes appearing in the tree and the number of total features, in parentheses. The P-A tree, for example has six features from three classes. The last row summarizes the number of test items that relied on a feature class and the number of features from that class across all five test items, in parenthesis. For example, Rate and Length were present in every test item and lexical sophistication was present in all but one test item. The table suggests that across all test items there was good coverage of feature classes but length was especially well represented. This is to be expected with a group heterogeneous in speaking proficiency. The length features often were used to classify students in the lower scores, that is, students who could not manage to speak sufficiently to be responsive to the test item.

9 Discussion

9.1 Speech recognition

We successfully adapted an off-the-shelf speech recognition engine for the purpose of assessing spontaneous speaking proficiency. By acoustic and language model adaptation, we were able to markedly increase our speech recognition engine's word accuracy, from initially 15% to eventually 33%. Although a 33% recognition rate is not high by current standards, the hurdles to higher recognition are significant, including the fact that the recognizer's acoustic model was originally trained on quite different data, and the fact that our data is based on highly accented speech from non-native speakers of English of a range of proficiencies, which are harder to recognize than native speakers.

9.2 SVM and CART models

Our goal in this research has been to develop models for automatically scoring communicative competence in non-native speakers of English. The approach we took is to compute features from ASR output that may eventually serve as indicators of communicative competence. We evaluated those features (a) in quantitative respect by using SVM models for score prediction and (b) in qualitative respect in terms of their roles in assigning scores based on a human criterion by means of CART analyses.

We found in the analysis of the SVM models that despite low word accuracy, with ASR recognition as a basis for training and testing, scores near inter-rater agreement levels can be reached for those items that include a listening or reading passage. When simulating perfect word accuracy (in the align-align configuration), 4 of 5 test items achieve scoring accuracies above the mode baseline. These results are very encouraging in the light that we are continuing to add features to the models on various levels of speech proficiency.

Test item	Length	Lexical sophistication	Fluency	Rate	Content	Total: # classes (# features)
P-A	4	1	0	1	0	3 (6)
P-C	4	0	1	1	1	4 (7)
P-T	2	1	0	1	1	4 (5)
P-E	1	1	2	1	1	5 (6)
P-W	1	2	0	1	0	3 (4)
Total # classes (# features)	5 (12)	4 (5)	2 (3)	5 (5)	3 (3)	19 (28)

Table 3: Distribution of features from the nodes of five CART trees (rows) into feature classes (columns). The "totals" in the last colunmn and row count first the number of classes with at least one feature and then sums the features (in parentheses).

CART trees have the advantage of being inspectable and interpretable (unlike, e.g., neural nets or support vector machines with non-linear kernels). It is easy to trace a path from the root of the tree to any leaf node and record the final decisions made along the way. We looked at the distribution of features in these CART tree nodes (Table 3) and found that all the different categories of features were used by the set of trees. For all 5 test items, most classes occurred in the nodes of the respective CART trees (with a minimum of 3 out of 5 classes).

10 Conclusions and future work

This paper is concerned with explorations into scoring spoken language test items of non-native speakers of English. We demonstrated that an extended feature set comprising features related to length, lexical sophistication, fluency, rate and content could be used to predict human scores in SVM models and to illuminate their distribution into five different classes by means of a CART analysis.

An important step for future work will be to train the acoustic and language models of the speech recognizer directly from our corpus; we are additionally planning to use automatic speaker adaptation and to evaluate its benefits. Furthermore we are aware that, maybe with the exception of the classes related to fluency, rate and length, our feature set is as of yet quite rudimentary and will need significant expansion in order to obtain a broader coverage of communicative competence.

In summary, future work will focus on improving speech recognition, and on significantly extending the feature sets in different categories. The eventual goal is to have a well-balanced multi-component scoring system which can both rate non-native speech as closely as possible according to communicative criteria, as well as provide useful feedback for the language learner.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bernstein, J. (1999). *PhonePass Testing: Structure and Construct*. Menlo Park, CA: Ordinate Corporation.

Bernstein, J., DeJong, J., Pisoni, D., & Townshend, B. (2000). *Two experiments in automatic scoring of spoken language proficiency*. Paper presented at the InSTIL2000, Dundee, Scotland.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Belmont, California: Wadsworth Int. Group.

Burger, S. (1995). *Konventionslexikon zur Transliteration von Spontansprache*. Munich, Germany.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47.

Cucchiarini, C., Strik, H., & Boves, L. (1997a, September). *Using speech recognition technology to assess foreign speakers' pronunciation of Dutch.* Paper presented at the Third international symposium on the acquisition of second language speech: NEW SOUNDS 97, Klagenfurt, Austria.

Cucchiarini, C., Strik, S., & Boves, L. (1997b). *Automatic evaluation of Dutch pronunciation by using speech recognition technology.* Paper presented at the IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000). *The SRI EduSpeak system: Recognition and pronunciation scoring for language learning.* Paper presented at the InSTiLL-2000 (Intelligent Speech Technology in Language Learning), Dundee, Scotland.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269-293). Harmondsworth, Middlesex: Penguin.

Meteer, M., & al., e. (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus*.Unpublished manuscript.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. New York, NY: Peter Lang.

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Development, 7*(26).

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*: Springer.

Xi, X., Zechner, K., & Bejar, I. (2006, April). *Extracting meaningful speech features to support diagnostic feedback: an ECD approach to automated scoring.* Paper presented at the NCME, San Francisco, CA.