2008

The Prosody of Uncertainty for Spoken Dialogue Intelligent Tutoring Systems

Bevan Jones
San Jose State University

The Prosody of Uncertainty for Spoken Dialogue Intelligent Tutoring Systems

A Writing Project
Presented to
The Faculty of the Department of Computer Science
San José State University

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science

by
Bevan K. Jones
May 2008

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

Dr. David Scot Taylor

Assistant Professor of Computer Science, San Jose Sate University, San Jose, California

---

Dr. Jeff Smith

Associate Professor of Computer Science, San Jose State University, San Jose, California

---

Dr. Elizabeth Owen Bratt

Senior Research Scientist of the Computational Semantics Lab, CSLI, Stanford University, Stanford, California

---

APPROVED FOR THE UNIVERSITY

Abstract

The speech medium is more than an audio conveyance of word strings. It contains meta information about the content of the speech. The prosody of speech, pauses and intonation, adds an extra dimension of diagnostic information about the quality of a speaker's answers, suggesting an important avenue of research for spoken dialogue tutoring systems. Tutoring systems that are sensitive to such cues may employ different tutoring strategies based on detected student uncertainty, and they may be able to perform more precise assessment of the area of student difficulty. However, properly identifying the cues can be challenging, typically requiring thousands of hand labeled utterances for training in machine learning. This study proposes and explores means of exploiting alternate automatically generated information, utterance correctness and the amount of practice a student has had, as indicators of student uncertainty. It finds correlations with various prosodic features and these automatic indicators and compares the result with a small set of annotated utterances, and finally demonstrates a Bayesian classifier based on correctness scores as class labels.

Acknowledgments

I owe much to the generous and patient tutelage of Liz Bratt, whose invaluable advice and assistance, in all aspects of this project have proven invaluable, both to this culminating experience of my computer science graduate study and to my general education and current direction in life, including my decision to go on in my studies in the disciplines of Cognitive Science and Linguistics. Without our frequent brainstorming sessions this thesis could never have existed. Indeed, starting out, I knew practically nothing of the scientific study of prosody and the tools for its pursuit. Also, I am grateful for her numerous reviews of the paper comprising much of the earlier work presented here, without which things would certainly be in a much sorrier state.

I would also like especially to acknowledge Professor Taylor for his kind and candid advice and encouragement. Without his encouragement to pursue my own interests, no matter how far afield they may lead, this entire undertaking would likely have been completed much earlier -- but would undoubtedly have been considerably poorer in content and I less rich for the experience. And, of course, he also has had a definite impact on my current professional course.

There are, naturally, many others to which I owe much. However, rather than list them all here in a hopelessly doomed effort to somehow do justice to all they have done for me, I shall simply endeavor to repay my numerous debts in the best way I know how. I will simply exert myself in putting all their contributions to the best use I can, and I will make every effort to afford others the opportunity to do the same.

Table of Contents

INTRODUCTION

Objectives & Summary of Contributions

This work has both immediate and long term objectives.

The primary objective is to construct a classifier for spoken language that can assess the speaker's degree of confidence about the content of his speech. The first portion of the work is, thus, directed at identifying the features of speech that are most likely to be helpful for such a classifier. The investigative tools and techniques employed for this first portion are consistent with traditional statistical exploratory analysis and hypothesis testing. The later portion then puts these finding to work by implementing a Bayesian classifier that demonstrates to what extent these statistically significant features can be relied upon to classify utterances.

The longer term objective deals with improving the effectiveness of educational software, and finding ways that a speaker confidence classifier can be put to use. More specifically, this longer term objective involves identifying applications for the classifier in assessing student performance and tailoring automatically delivered educational material more closely to student competency and emotional state. Thus, while the majority of the work deals most directly with the business of the analysis of language and the design of a classifier, much discussion contained here is dedicated directly to the larger objective of improving the state of the art in educational software. However, while applications are discussed, the actual integration of the classifier into a live system falls outside the scope of this project.

The main innovations of the work involve the use of automatically verified measures of speaker confidence for both statistical analysis and machine learning, with secondary innovations in tailoring a Bayesian classifier to the particular problem. More specifically, while prior work has already identified the prosodic patterns of various phenomena related to speaker confidence, this previous work has generally required masses of hand labeled data. This work first verifies that the prosodic features that were relevant for the hand labeled data remain significant for mechanically generated labels, and then proceeds to build and assess the effectiveness of a classifier based on these features.

Motivation & Background

One-on-one dialogue between student and tutor affords particular opportunities and advantages for enhancing student learning beyond more traditional classroom activities,

sometimes by as much as two standard deviations (Bloom, 1984). Typically, however, resources limit a teacher's ability to devote one-on-one time for adapting material directly to individual student needs, and computers may be employed. Educational software incorporating student modeling techniques (Conati et al., 2002; Rosé et al., 2001; Evans & Michael 2006) can adapt material to student performance over the course of a session, using a technique known as macro-adaptation (Shute, 1993). This allows a system to home in on and address more directly individual student needs within a single session. Furthermore, if a spoken dialogue interface is employed, still more opportunities arise.

Some intelligent tutoring systems such as Andes employ a strictly graphical interface (Conati et al., 2002). However, a speech interface may sometimes offer a more natural method of delivering learning material and for testing its assimilation (Pon-Barry et al., 2004; Litman & Silliman, 2004). In such cases, student speech may contain far more information than would normally be discernible from the literal transcription, possibly containing information on the student's emotional or cognitive state (Ang et al., 2002; Berthold & Jameson, 1999). Specifically, it may be possible to detect the relative confidence or uncertainty of the student in delivering a given utterance (Forbes-Riley & Litman, 2007; Pon-Barry et al., 2006). The addition of this dimension to tutoring system input offers opportunities for enhancements in at least two areas.

- **Student performance assessment**: It may be possible to exploit the additional information for greater precision in assessing the area of student difficulty. Students may signal uncertainty localized to specific items within an utterance by, for instance, pausing just prior to the item or terminating it with a rising tone. Furthermore, even if the location of the cue does not clearly mark the source of confusion, the system can ask questions designed for homing in on it. In such cases, even when a student's response is perfectly correct, the additional information about student uncertainty may cue an instructor that the student may benefit from further instruction in this area.

- **Selection of appropriate tutoring tactics**: Besides pinpointing the exact item of difficulty, such information also serves in assessing the student's emotional state. This has important implications for the appropriateness of a given pedagogical tactic, as tactics that are appropriate for less confident students may not be for more confident students and vice versa. Tactics such as model, scaffold, fade (Collins et al., 1989) rely implicitly on such assessments. Moreover, an automated tutor can be designed to respond in such a way as to increase student persistence in the face of waning confidence (Aist et al., 2002). It is worth noting that human tutors have been observed to give different responses based on student uncertainty (Forbes-Riley & Litman, 2007).

Thus, a tutoring system capable of identifying uncertainty in the student's speech could give feedback, offer hints, ask leading questions, prompt for further explanation, or employ other pedagogical tactics for identifying the source of student confusion and resolving it.

This paper pursues this avenue of investigation, examining audio data recorded in the course of experiments with Voice-Enabled DCTrain (Peters et al., 2004), an intelligent tutoring and spoken dialogue system designed for training US Navy personnel as Damage Control Assistants. This work attempts to answer the question of whether this system or other similar systems might be enhanced to exploit prosody for adapting to student state. The goal for this work is to highlight automatically extractable features that interact closely with speaker uncertainty, thus facilitating machine learning approaches to automatic classification.

The direct modeling approach (Shriberg & Stolcke, 2004) to this would involve annotating utterances with a listener's assessment of speaker emotional state (e.g. questioning, hesitant, neutral), followed by an application of machine learning techniques to automatically map these high-level annotations to prosodic features that can be extracted automatically from utterance audio recordings. However, it can be difficult for listeners to identify emotional state, and the number of very clear examples may be so few as to lack a representative sample, leading to overfitting in any machine learning approach (Tan et al., 2006). Furthermore, generating such annotations is both expensive and error prone. Any means of automating or semi-automating the process of generating or verifying the annotations would be valuable.

The problem can be explored via two alternative automatically generated proxy measures of uncertainty. First, experience in doing a task usually leads to increased confidence. Thus, the number of times a student has performed a task can provide some information about whether the student's utterances are likely to be confident or not. Later, this claim is substantiated with the Voice-Enabled DCTrain corpus by showing that students do, in fact, improve with practice. Second, increased competence leads to increased confidence. More specifically, a student's utterances should exhibit greater confidence when the student has mastered the material sufficiently to consistently score well, a phenomenon that has been observed in related work (Forbes-Riley & Litman, 2007). Of course, features that correlate with both the amount of practice a student has had and the correctness of his utterances are even more likely to be useful indicators of uncertainty.

The literature suggests six features as likely candidates for automatically extractable confidence indicators: pause rate, speech rate, pitch rise, mean pitch, mean intensity, and the amount of speech production. Pause rate, speech rate, the change in pitch at the termination of phrases and utterances, and the mean intensity of utterances have all been

used in automatically classifying dialogue acts (Shriberg et al., 1998), and pitch rise and pause rate, in particular, were both found to be highly useful in classifying yes-no questions, which may bear a resemblance to uncertain statements. Furthermore, pitch, intensity, and speech rate have been used in distinguishing uncertainty from confidence and frustration (Zhang et al., 2003). In addition, pause rate and speech rate are useful indicators of cognitive load (Berthold & Jameson, 1999; Clark & Fox Tree, 2002). Because cognitive load speaks directly to the difficulty of the student's task, it seems reasonable that easier tasks would be more likely to induce confidence in speakers. Finally, Core et. al. (2003) demonstrate that the amount of speech production correlates with student learning gains, and so this can be considered another feature potentially relating to confidence.

Subsequent sections define these features in detail; describe the corpus to be used; describe the approach and methods for finding prosodic patterns of uncertainty; discuss the correlations found in the data; discuss the methods and the relationship between correctness, practice, and confidence; discuss the architecture of the naïve Bayesian classifier; evaluate the performance of the classifier; and conclude with a summary of the contributions made with this work and suggestions for further exploration.

FEATURE DEFINITIONS

This section provides definitions of the automatically extractable features. For many of the features (pitch information, intensity and speech rate) phrase and utterance measurements are normalized by the student's mean values. This is generally in keeping with other work in prosodic analysis (Shriberg et al., 1998; Zhang et al., 2003), and it allows the capture of speakers' deviations from their own mean performance. Praat (Boersma & Weeninck. 1996) can then be employed to extract pitch and energy information as per (Huang et al., 2006). Word alignment information is obtained via a preprocessing pass with the Sphinx 2v0.5 recognizer (Huang et al., 1993), using the "Communicator" acoustic models (Bennett & Rudnicky, 2002). Wherever the text of utterances is required, transcriptions rather than the Automatic Speech Recognition (ASR) hypotheses are used in order to get the clearest look at the prosodic phenomena, although a fully automated system could use the ASR hypothesis.[1]

Pause Rate: Here, pauses are defined as silences in duration exceeding some threshold between bounding words, as distinct from periods of silence at the beginning or ending of utterances. That is, each pair of adjacent words contributes one opportunity for a pause, and the duration of whatever silence there may be between words determines if a pause is, in fact, present. The silence threshold employed in the experiments for identifying

1 The Nuance recognizer has an average word error rate of 5.7% for this corpus.

pauses was set at 200 ms, as per (Müller et al., 2001; Berthold & Jameson, 1999). This threshold serves to cover potential measurement error in the alignment timings and may otherwise ensure a certain degree of significance to the pause. In addition, for the Voice-Enabled DCTrain corpus, silences in excess of 900 ms were used to automatically end-point utterances when the ASR recorded the wave files. Thus, all silences between words measured between 0.2 and 0.9 seconds in duration identify a pause. The pause rate of an utterance or phrase is defined as the total number of pauses measured per the number of pause opportunities. For example, an utterance of five words contains four pause opportunities. Thus, if one pause were observed, the pause rate for the entire utterance would be 0.25.

Speech Rate: To approximate speech rate, mean word durations are tracked for each student, eliminating any periods of silence from consideration. Then, to determine the speech rate of a given word sequence as spoken by a particular student, the duration of each word instance within the sequence is divided by the corresponding student's mean duration for that word. By this method, a normalized word duration is obtained for each word instance, and from this an average of these over all the word instances in the sequence to get a measure for the entire sequence. Hence, an average normalized word duration of 1.3 implies that words within the utterance were on average 30% longer in duration then normal for the given subject. The additional refinement is made by tracking the durations of phrase terminal words separately, so that phrase terminal words are normalized by the average duration of the particular word as it occurs in the terminal position. This resolves a potential problem where utterances containing more phrases may tend to longer average word durations simply because of the well known tendency in English to prolong phrase and utterance final words (Wightman et al., 1992). This normalized word duration feature, where terminal words are normalized by their mean terminal position and all other words are normalized by their non-terminal duration, is referred to as word_dur_norm_avg_phrase_aware. The normalized word durations can then be converted into the reciprocal normalized speech rate value:

speech_rate = 1 / word_dur_norm_avg_phrase_aware.

Pitch-Rise: The feature used to measure pitch rise is the relative change in average pitch between the last two 200 ms segments preceding a word ending. That is, as defined by (Shriberg et al, 1998),

rel_f0_diff = end_f0_mean / pen_f0_mean.

Here, end_f0_mean is the average pitch in the last 200 ms and pen_f0_mean is the average pitch in the penultimate 200 ms segment. Also in keeping with Shriberg et al. (1998), rel_f0_diff is normalized by the subject mean over all utterances to get the feature

refer to in this paper as rel_f0_diff_normal.

Mean Pitch: In addition to pitch rise, the mean f0 values of utterances are also examined, again normalized by subject mean across all utterances. This value is measured from the portions of utterances excluding silences, as periods of silence would skew f0 measurements toward zero and potentially confound mean f0 values with the proportion of pauses.

Intensity: To test the supposition that more confident utterances, spoken with greater authority, may have greater overall intensity, the energy of utterances and phrases is measured. The intensity of an utterance is defined as mean decibels normalized by the subject's average intensity over all utterances, intensity_db_normal. Note that this normalization serves two purposes. First, as with pitch and speech rate information, normalization allows a clearer focus on the primary interest of the speaker's deviation from their own average intensity. Also, and just as importantly, it accounts for utterance recording circumstances such as microphone distance, room acoustics, and volume settings. So as not to confound this measure with utterances containing long periods of silence, again, all periods of silence are excluded from the computed average.

Speech Production: Though not strictly prosody, the amount of speech production is a useful complement to prosody. Other work has found correlations between amount of speech production and student learning gains (Core et al., 2003), as students that produce more utterances and longer utterances generally learn more during tutoring sessions, and the possibility of a similar relationship with student confidence is explored here. It may be that the student, sensing his mastery of the material, gains confidence, and this increasing confidence leads to more speech production. Specifically, we may expect that more confident students may issue longer and more ambitious utterances. Thus, a look at utterance lengths may be enlightening. This can generally be measured in terms of number of words per utterance. Alternatively, it may be profitable to examine the number of phrases contained within an utterance.

Intuitively, one may expect pauses, word-durations, and pitch-rises to decrease with student confidence, while intensity and the amount of speech production should increase. These features are considered for entire utterances as well as localized to particular words or phrases. Tying features to particular phrases or words allows the test of the claim that prosody can be exploited for greater precision in identifying areas of student difficulty.

THE VOICE-ENABLED DCTRAIN CORPUS

DCTrain is a training system designed to simulate realistic conditions aboard a US Navy

ship for training Navy personnel (Damage Control Assistants) in coordinating ship damage control (Bulitko & Wilkins, 1999). To more closely approximate the true-to-life spoken command style, DCTrain was retrofitted with a speech interface (Peters et al., 2004). The research documented here is built on data collected in the course of three different experiments with Voice-Enabled DCTrain. The first two took place during 2004 with subjects drawn from the Stanford University student population. The third was conducted in 2005 with students from the US Naval Academy, Annapolis.

The entire corpus consists of 283 subjects and 17,129[2] utterances. The 252 subjects and 3,483 utterances judged for correctness are a subset of the corpus.

Table 1: Corpus Summary Statistics

	Spring 2004 - Stanford	Summer 2004 - Stanford	Winter 2005 - USNA	Total
Words	23822	29400	46050	99272
Utterances	4503	5504	7122	17129
Corrected Utterances	716	1129	1638	3483
Subjects	33	44	205	283
Corrected Subjects	32	43	177	252

Language Description and Correctness Ratings

The DCTrain simulator allows students to experience damage control scenarios on a US Naval ship, where the student plays the role of a damage control officer. The majority of student utterances consist of orders to repair teams in various areas of the ship and use a fairly specialized subset of English grammar and vocabulary. The result of this specialization is that word order and the number of words is held relatively constant, thereby reducing the number of variables necessary for consideration and facilitating a

2 These 17,129 utterances comprise roughly 60% of a larger corpus, and are those for which it was easiest to recover forced alignment information.

more detailed focus on prosody.

Utterances judged for correctness are of two types, as exemplified by the following two transcripts:

repair two investigate compartment two tac two two zero tac four tac alpha (1)
Repair Two, investigate compartment 2-220-4-A.

repair three dca set fire boundary primary forward two zero zero (2)
Repair Three, D.C.A., set fire boundary primary forward 200.

Commands of type (1) identify a repair team and compartment, while those of type (2) may or may not identify a repair team but always identify the boundary to set in order to contain a crisis such as a fire or flood. The information critical to the correctness scores are underlined for emphasis. The correctness of type (1) utterances depends upon whether the specified compartment falls within the jurisdiction of the specified repair team. In this particular instance, the student addresses repair team two, ordering them to investigate the compartment with designator "2-220-4-A". Note the use of the US Navy alphabet letter "alpha" for "A". The correctness of utterances of type (2), on the other hand, depends on the appropriateness of the boundary to the location of the crisis. In this instance, the student has ordered repair team three to set a boundary against fire spread, with designator "primary forward 200." Whether this is the correct boundary or not depends on whether it is either immediately (or one boundary removed) aft or fore of the compartment or compartments containing the crisis. These two particular utterances types are singled out simply because they can be judged for correctness independent of dialogue context. That is, they are self-contained, each containing all the information necessary for assessing correctness.

It is important to note that these two utterance types constitute a sizable but definite minority of utterances, about 20% of the corpus. There are other actions that the student may take which DCTrain scores but that no attempt is made here to score. Furthermore, as demonstrated in (2) it is common for repair team addresses to appear in utterances without accompanying compartment designator. Thus, by limiting investigations to utterances with both compartment and repair team, correctness is gathered for only some 31% of the utterances containing repair teams. Also, the measure used in this work for boundary phrase correctness is less precise than the one DCTrain uses, since there can be multiple compartments involved in a scenario as a whole, while only some subset of these are active at any given moment, and each boundary is checked against the union of correct boundary sets for all compartments active in a given experiment, while DCTrain checks only against the potentially much more precise set of boundaries for compartments currently active at the time of the order. As a matter of convenience, because the correctness information was not readily transferable from DCTrain logs to

the corpus of utterances, this approximation was used as a starting point[3].

METHODOLOGY: PRACTICE, CORRECTNESS, AND CONFIDENCE

The corpus contains three different human-annotated labels for uncertainty: hesitant, question-rise, and uncertain. The annotators were allowed to use any of these labels freely and were not required to identify confident utterances. T-tests find highly significant differences in mean pause rate and normalized word durations for hesitant utterances as compared to other utterances (at the $p < 0.001$ level for each), where hesitant speech contains both more pauses and words of longer duration (Figure 1). For utterances marked as containing a question-rise, a higher mean value is found for the pitch-rise feature, rel_f0_diff_normal ($p < 0.001$), as shown in Figure 2. However, the corpus contains only some 23 "hesitant" annotated utterances, 21 "question-rise" annotated utterances, and 67 "unsure" annotations, totaling 111 out of 17,129 utterances. Small sample size leads both to larger variances and reduced accuracy in estimating population statistics. Thus, the smallness of the sample of uncertain utterances not only makes it unlikely that the sample represents the full range of behaviors but also generally reduces

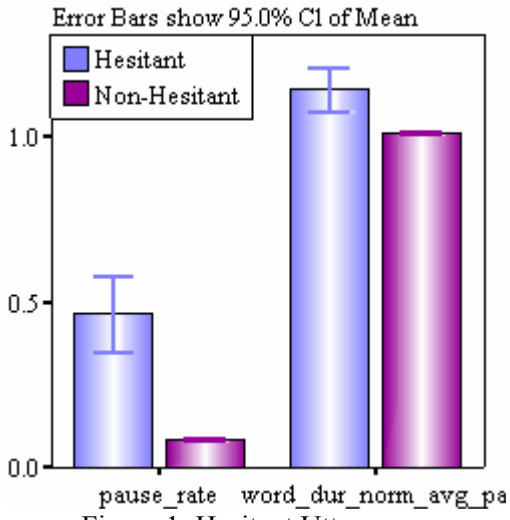

Figure 1: Hesitant Utterances

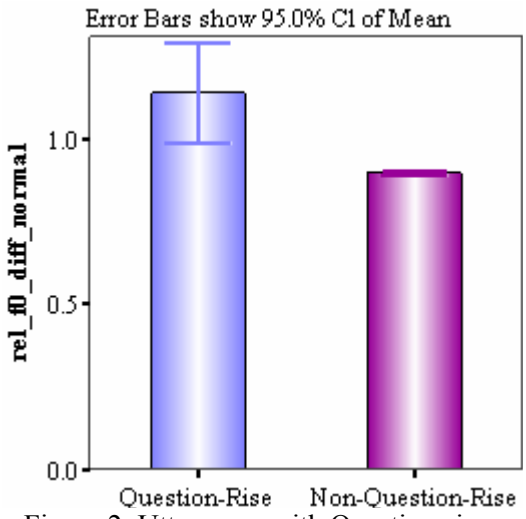

Figure 2: Utterances with Question-rise

the accuracy of statistical measurement. Nevertheless, this sample does provide some support for the hypotheses.

3 DCTrain has its own representation of the correctness of utterances, but there are potential problems in relying on logs for this info. First, speech recognition errors may mean DCTrain is not interpreting the utterance correctly. Second, DCTrain allows a command to be built up over the course of several dialogue turns, filling in missing parameters each turn. It only assesses correctness when the command is complete. The nature of prosody in multi-turn commands could easily be quite different. Third, DCTrain could not completely solve boundary correctness because if several fires are burning, DCTrain only matches boundary commands to its set of needed boundaries, and does not represent which boundaries are intended for which fire.

It seems, however, that there may be many more examples of uncertainty in the corpus than the annotations alone indicate. Any annotations for uncertainty must necessarily be based on annotator perception, and it can be difficult for annotators to correctly identify the actual state of the subject.

One important source of difficulty resides in the fact that some of the prosodic features considered for classifying uncertainty may be voluntary, and, thus, may not always coincide with the speaker's confidence state. Filled pauses, for instance, used to announce forthcoming speech delays, possibly due to cognitive load, are one example of such voluntary expressions of speaker self-assessment (Clark & Fox Tree, 2002). Speaker choice in issuing such signals is governed by many factors, one of which is the perceived nature and role of the interlocutor (Shechtman & Horowitz, 2003; Reeves & Nass, 1996). If the subject perceives the system strictly as a machine, voluntary expression of uncertainty may not be present at all (Shechtman & Horowitz, 2003). Alternately, according to the argument of Reeves and Nass, even if the students were to interact with the system as they would with humans, the perceived social role assumed by the system would shape student interactions accordingly. In this case, the simulated entities in the damage control scenario are mostly subordinate officers, and student utterances are less likely to contain signals of uncertainty.

While these factors make it difficult for annotators to diagnose uncertainty they do not indicate whether or not students themselves are confident. That is, there may yet be many instances of uncertainty that are not reflected in the annotations, and, owing to the difficulty of the diagnosis, there may even be cases of misidentified uncertainty. Therefore, the annotations are supplemented by other measurable phenomena that serve as proxy measures of uncertainty.

Correctness & Confidence

One means of approximating uncertainty annotations is to rely on correctness scores, which in the Voice-Enabled DCTrain corpus are automatically scored. Research has shown that, while correctness is not identical with confidence, the two are closely tied (Forbes-Riley & Litman, 2007). Intuitively, this makes sense, in that students should be confident of successfully completing easier tasks, and easier tasks will also generally receive higher correctness scores. Nevertheless, students may be mistaken about what they know, leading to overconfidence or under-confidence, and they may make correct guesses that outstrip their true understanding. That is, students may be correct yet uncertain, or they may be incorrect but still confident. These issues introduce noise into the approximation, but the greater numbers of utterances with scores are a helpful addition to the study of the prosody of uncertainty. Thus, t-tests are performed for measuring the significance of mean differences between correct and incorrect utterances/phrases to parallel tests for confidence vs. uncertainty.

Increasing Confidence with Practice

Another automatically measurable indicator of student confidence is the amount of practice a student has had. The time the subject has spent in practice with the tutor is measured by the number of utterances or phrases the student has produced just prior to and including the utterance under consideration. This chronological ranking allows averaging across subjects for each given utterance/phrase number, potentially revealing trends over time. Then, to supply some quantitative measure of the strength of any such trends, p-values are computed for two-tailed Pearson correlation for linear regression. Relationships are not expected to necessarily be linear, but the correlations can indicate generally increasing or decreasing trends.

The idea that confidence increases with utterance or phrase number is an assumption that merits closer inspection[4]. While it seems natural to assume most students will gradually become more comfortable, we first examine the relationship between chronological ranking of utterances and other factors that may either contribute to or be the result of improving student confidence. Specifically, relationships are sought with regard to disfluencies such as word fragments and broken-off utterances, phenomena that have been commonly cited as relating to cognitive load (Berthold & Jameson, 1999). Generally, student confidence in some sense speaks to the ease or difficulty of the task at hand, and cognitive load is a closely related factor. Similarly, a search is conducted for prosodic features that may occur in conjunction with correct or incorrect utterances, another indicator of the general difficulty of the task. We also take a look at the other less directly related phenomena of speech recognition problems (word error rate and rejection rate).

Other work has found correlations between recognition problems and unusually fast or slow speech, out of vocabulary words, and speaker self-repairs related to disfluencies (Shinozaki & Furui, 2001; and Shinozaki & Furui, 2002; Hirschberg et al., 2004). Furthermore, others have observed relationships between the heightened emotional state of the speaker (such as frustration) and speech recognition problems (Rotaru & Litman, 2006). While frustration may or may not relate to disfluencies, its presence or absence speaks to the smoothness of student interaction with the system. In general, speech recognition performance seems closely related to the quality of speech production and to

4 The reasonableness of the assumption depends on the dynamics of the interaction between the system and the student. It has been observed that users can enter into negative cycles of interaction, where recognition failures, for instance, may prompt the user to alter his speaking manner in such a way that it may be even harder to automatically recognize, resulting in even worse recognition performance (Soltau & Waibel, 1998). In such a scenario, confidence may very well never come to be the dominant factor in the user's experience. However, negative dynamics of this sort would be expected to exhibit symptoms such as increasing disfluencies, poorer recognition performance, and generally poorer student performance. Thus, we first look at these symptoms, including other signs of improvement or problems, such as cognitive load, correctness ratings, and utterance lengths.

the general smoothness of the interaction between the speaker and the system. Thus, declining rejection counts and Word Error Rate may point to a general improvement in user experience and an increasing sense of confidence. The data examined here supports these findings (Figures 3 and 4), with significant differences between the mean word error rate for hesitant and non-hesitant utterances ($p<0.001$) and the mean number of recognition failures for utterances marked with a question-rise as compared to all other utterances ($p<0.005$), as computed via t-tests. In both cases, the uncertainty marked utterance was much more likely to be associated with a speech recognition problem.

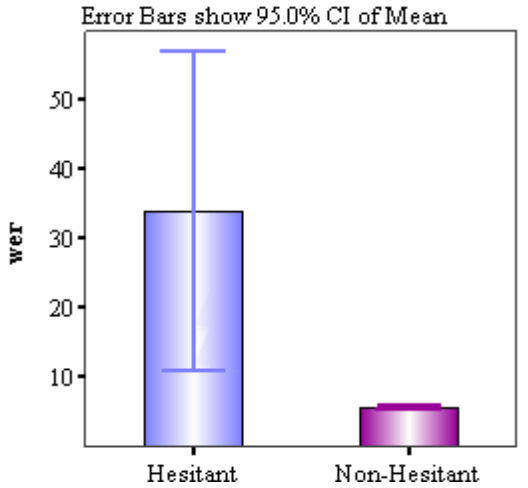

Figure 3: ASR Problems & Hesitancy

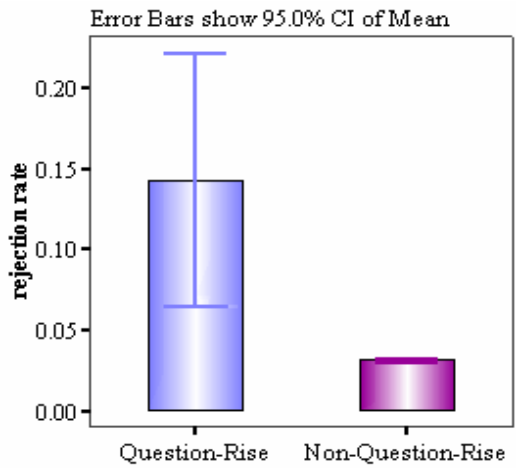

Figure 4: ASR Problems & Question-rise

In fact, averaged over all subjects in the data, it is observed that disfluencies and recognition problems decrease over time (Figure 5). Furthermore, student responses to simulated casualties are more consistently correct, for both boundary phrases and repair team addresses (Figure 6).

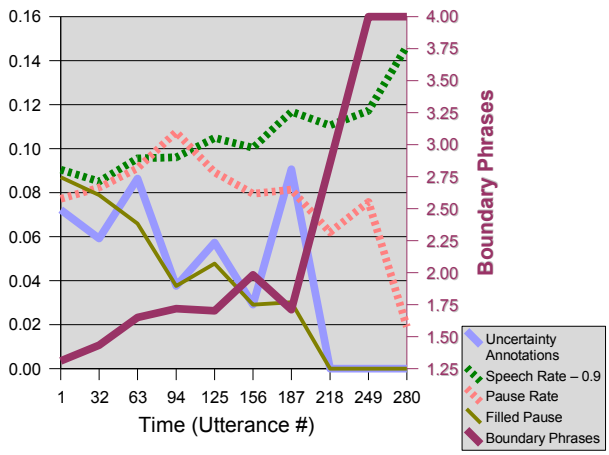

Figure 5: Confidence Indicators vs. Time

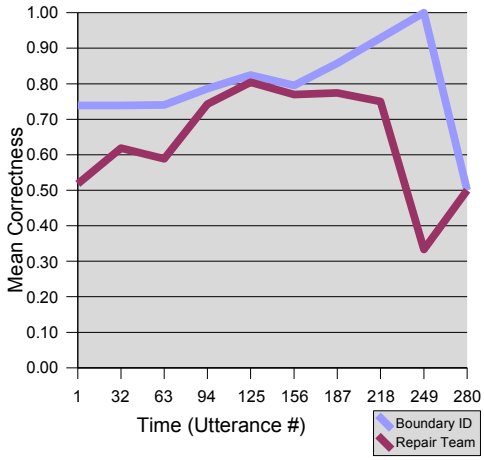

Figure 6: Correctness vs. Time

From (Figure 7) it can also be observed that the student's rate of speaking increases (word

durations get shorter over time) while pauses (marked or unmarked by "uh" or "um") grow less frequent, both phenomena cited as signs of decreasing cognitive load (Berthold & Jameson, 1999). Moreover, there is a significant increase in the number of boundary phrases students typically incorporate into a single utterance. This kind of increase in speech production has been noted to often accompany learning gains in student interactions with tutors (Core et al, 2003), and seems likely to correspond to increasing student confidence. Furthermore, we observe the uncertainty annotations decreasing in frequency over time. All of these combined factors argue strongly that student confidence does indeed increase with time spent practicing with the system.

Note that most of the disfluency measures presented are automatically detectable. They include the rate at which a subject tends to break off in the middle of boundary phrases, repair teams, and compartment phrases. Broken-off utterances can be identified by looking at the constituent phrases and matching phrase beginnings to the following string of words. If phrase terminating words are not matched, the utterance is marked as having been broken off.[5] Rejections occur as a result of very low acoustic likelihood measures, as judged by the speech recognizer. Word Error Rate (WER), on the other hand, requires a gold standard transcription for comparison, but these scores can be crudely approximated by confidence scores from the speech recognizer. Much better than this, however, is to employ prosody to improve on the simple acoustic likelihood (Hirschberg et al., 2004). Using this approach to approximate WER, only word fragments cannot be easily automatically detected.

5 This method seemed sufficient for our purposes. While it allows for utterances that are not necessarily "broken-off" to be labeled as such, these cases were generally due to the presence of some other disfluency type.

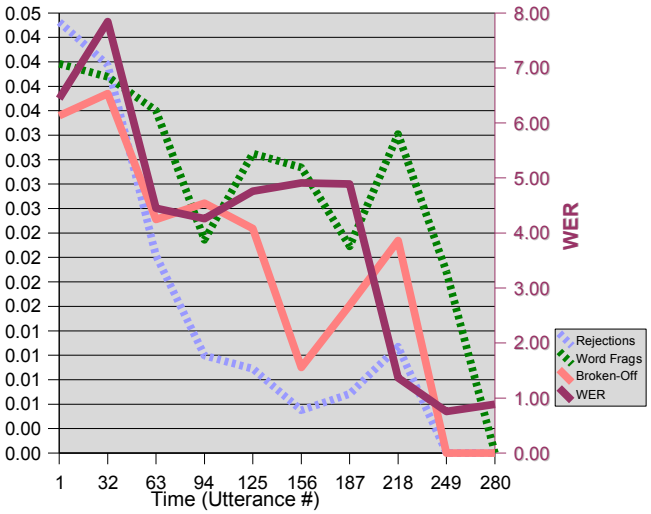

Figure 7: Disfluencies vs. Time

Furthermore, the confidence indicators, including correctness ratings but excluding the annotations themselves, are all automatically detected. So, these features can be extended from this statistical analysis to an automatic classifier that can then be incorporated into a tutoring system.

RESULTS

Amount of Production

Core et al. (2003) show that student dialogue contributions as measured in words closely relate to learning gains. The data bears these findings out, with regard to the number of words per utterance, but the relationship manifests differently for the two different utterance types at the focus on of this work. Specifically, it is found that utterances containing boundary phrases grow in length with practice and experience, while utterances containing repair-team/compartment pairs grow shorter. The density of the student's delivery of the necessary information seems to be the more fundamental measure of student competence than raw word counts.

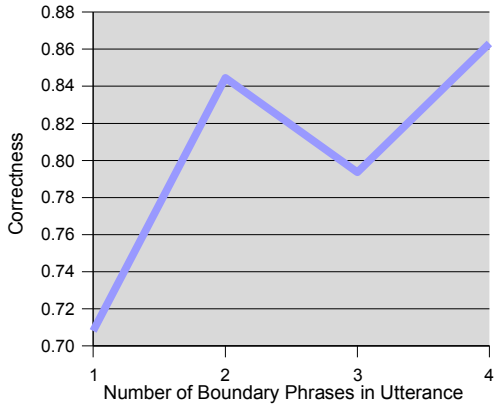

Figure 8: Number of Boundary Phrases vs. Correctness

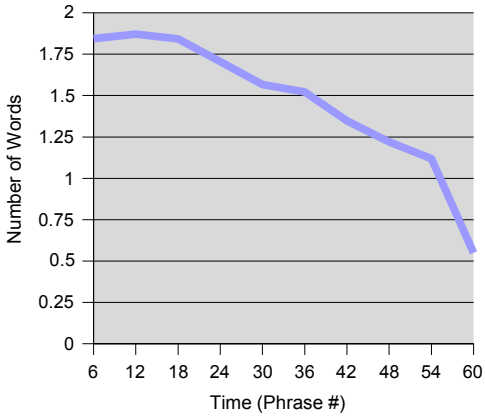

Figure 9: Optional Words in Repair Team Address vs. Time

When a casualty such as a fire occurs, one of the steps in controlling the fire is to set boundaries to contain the damage. There are four boundaries for each compartment: two aft-ward and two forward of the compartment. The student is to set all four boundaries but is permitted to specify them in any order, and any number at a time. That is, students may specify all four in the same utterance, or break them up into smaller sets over successive utterances. Thus, there is some variability in the number of utterances that seems related to the student's familiarity with the task and general sense of confidence. We observe that the number of boundary phrases within an utterance increases with practice ($p < 0.001$) and with utterance correctness ($p < 0.001$), as shown in Figure 8. These correlations strongly suggest phrase counts as a measure of confidence, and this paper refers to the count of information items per utterance as the "informativeness" of the utterance, where the more informative an utterance is, the more confident it appears.

Since repair-team/compartment utterances are more constrained in the number of information items that can be delivered per utterance, just one pair per utterance, the related concept of "conciseness" is found to be of greater utility. Here, the conciseness of an utterance is defined as the fewness of words used to deliver the necessary information. Thus, while there is no observable relationship with correctness, one may see from Figure 9 a significant decrease in the number of words in repair team addresses with practice ($p < 0.001$). Repair team addresses may consist of simply two words such as "repair three" or, at the other extreme, they may consist of as many as six words as in "net eighty to repair team three," where the "net eighty to" and the "team" are two different optional additions that students drop over time. Thus, while boundary utterances grow in length on average, repair team/compartment utterances actually shorten. That is, utterances get more dense and information rich with experience.

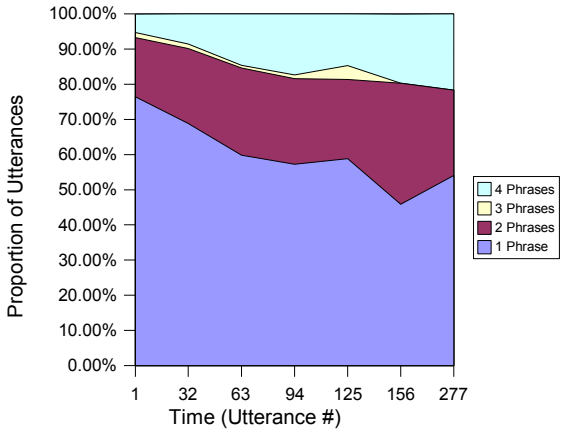

Figure 10: boundary Phrases in Utterances vs. Time

Students also demonstrate increasing competence by the manner in which they organize the information. Figure 10 shows that students tend to organize boundary phrases into sets of one, two, or four, with single-phrase utterances gradually decreasing in relative frequency over time even as the relative frequency of two- and four-phrase utterances increases. The physical layout of boundaries symmetrically about the compartment suggests a natural segmentation, and students gravitate toward this organization as they gain experience. On the other hand, we observe that three-phrase utterances remain relatively rare throughout, at about 1% of all boundary utterances. Furthermore, while informativeness still has an effect, since three phrase utterances are generally more correct than single phrase utterances, they are less likely to be correct than either two or four phrase utterances. Thus, the relationship between the number of boundary phrases in the utterance and student competence is not a linear one, but the logical organization of the utterance can be used in combination with phrase counts for greater accuracy in assessing confidence state.

Pauses

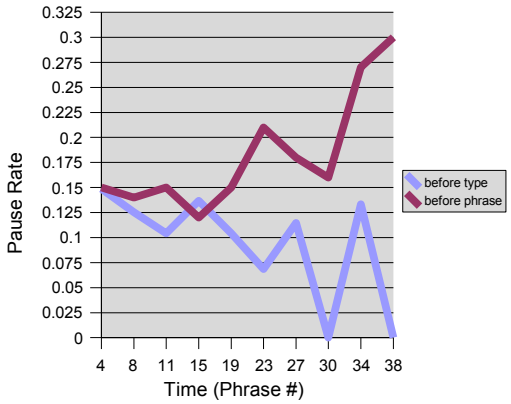

Figure 11: Compartment Phrase Pauses vs. Time

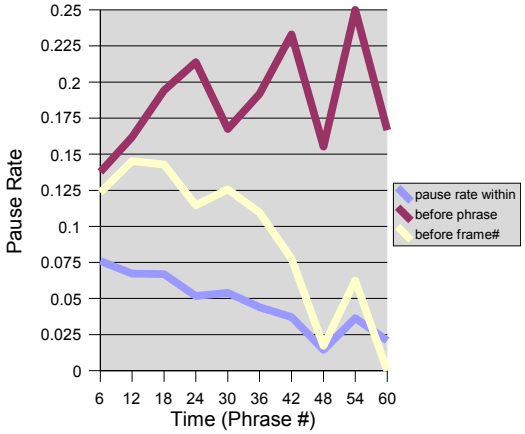

Figure 12: Boundary Phrase Pauses vs. Time

Intuitively, one expects more pauses to indicate less confident speech, and this intuition is supported both by the data and the literature on cognitive load, with one critical difference: while pauses within phrases do seem to decrease with confidence, pauses before phrases have the opposite relationship. That is, as depicted in Figures 12 and 13, pauses within boundary phrases decrease with practice ($p < 0.05$) and correctness ($p < 0.001$). However, students actually pause more with time at the grammatical points marking the beginning of boundary phrases ($p < 0.05$) and compartment phrases ($p < 0.01$), and more for correct boundary phrases ($p < 0.05$). The patterns correlated with correctness for boundary phrases are consistent with the observations for compartment phrases but with less significant trends (i.e., $p > 0.05$). Thus, it seems that ungrammatical pauses do in fact indicate problems, while grammatical pauses indicate either more careful planning of answers or greater fluency with the language.

We also observe from Figure 13 that pauses are much more frequent before critical sections of phrases, possibly allowing for the diagnosis of critical areas of difficulty. More specifically, while the difference between correct and incorrect utterances is less clear here, students still pause more frequently before the frame number of a boundary phrase identifier than elsewhere in the phrase ($p < 0.001$). These critical areas are explained by the fact that compartment boundaries are aligned with frame divisions within the ship, and the correct boundary can usually be ascertained directly from the frame number associated with the particular compartment. Thus, determining the frame number of the boundary is most of the task of determining the boundary in its entirety.

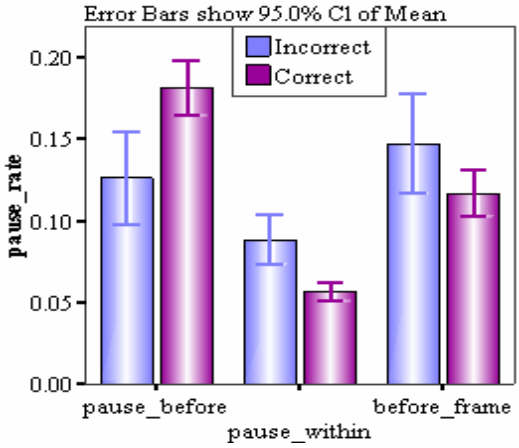

Figure 13: Pauses in Boundary Phrases vs. Correctness

Word Durations/Speech Rate

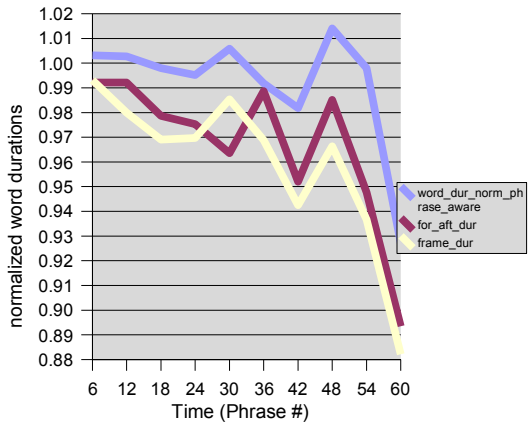

Figure 14: Boundary Phrase Word Durations vs. Time

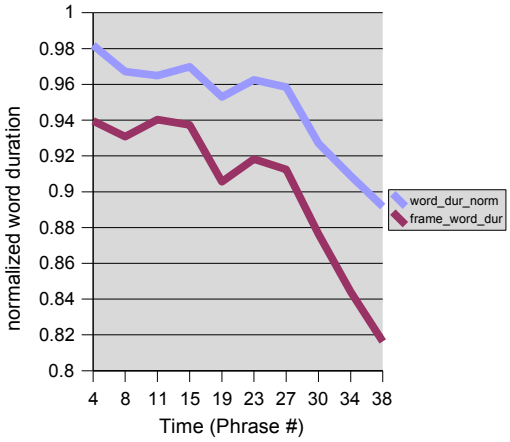

Figure 15: Compartment Phrase Word Duration vs. Time

Speech rate information is very similar to the pause rate information, where faster speech corresponds to greater competence. One difference, however, is that the duration of the key words of the boundary phrase are much more significant than the pauses preceding the key words, which were not found statistically significant. Over time, speech rate increases (see Figures 14 and 15) for both compartment ($p < 0.001$) and boundary phrase frame numbers ($p < 0.005$). Also, speech rate increases with correctness (Figure 16) for boundary phrase frame numbers ($p < 0.001$). Thus, word durations may serve better in pinpointing the exact place of difficulty within phrases. Otherwise, we observe essentially the same trends, where speech is generally faster (normalized word durations are shorter) within correct boundary phrases ($p < 0.001$) and speech rate also increases with practice

($p < 0.001$) for both boundary phrases and compartment phrases. It seems that speech rate is largely useful as a parallel measure of the phrase-internal pause rate.

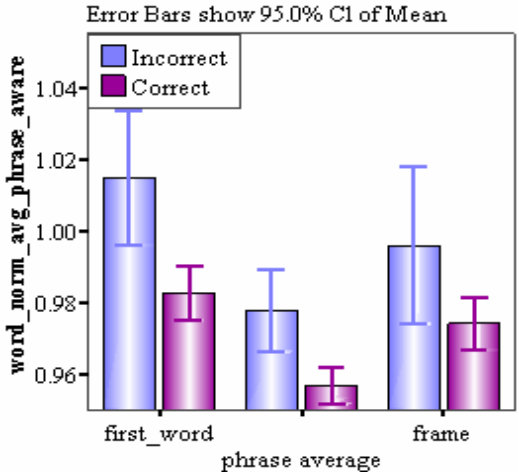

Figure 16: Boundary Phrase Word Durations vs. Correctness

Pitch Rise and Mean F0

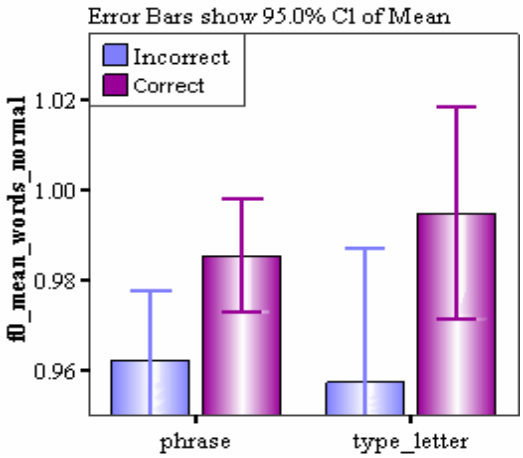

Figure 17: Compartment Phrase Mean F0 vs. Correctness

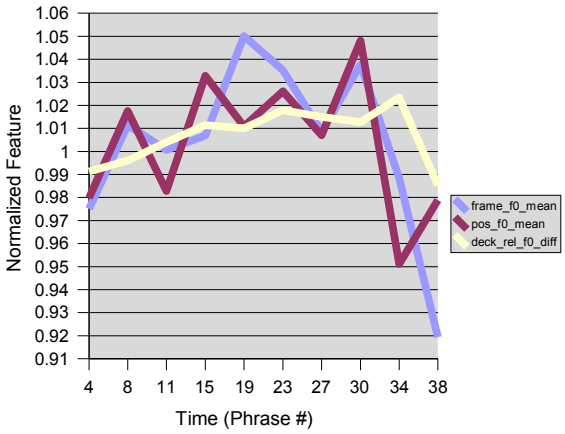

Figure 18: Compartment Phrase Mean F0 vs. Time

Interestingly, it is observed that within the corpus the much lower than average normalized pitch measures corresponds to the less confident phrases. Like speech rate, it offers some useful information for identifying critical areas of difficulty within utterances. Specifically, Figure 17 shows that the compartment usage type letter[6] has a

6. Note that the compartment letter is canonically expressed using the US Navy alphabet, and may require that the student either exercise his memory or consult a reference card to recall the correct word: "quebec" for "q", for instance.

closer to average normalized pitch for correct repair-team/compartment utterances ($p < 0.05$). Within compartment phrases in general (Figures 17 and 18), f0 rises with both correctness and time ($p < 0.05$ for both). It is perhaps worth noting that while most of the other features considered here are generally less significant for the compartment phrases than for boundary phrases; we observe the opposite effect with pitch information.

We can also observe, that the pitch rise feature, rel_f0_diff_normal, tends to fall more over time for the final word of repair team addresses ($p < 0.01$). However, somewhat less intuitively, rel_f0_diff_normal rises increasingly more often with time at the beginning of confident phrases ($p < 0.05$ for boundary phrases, $p < 0.01$ for compartment phrases), producing a higher overall mean f0.

Intensity

Like f0, an overall higher intensity within phrases seems to indicate greater confidence, as students speak more loudly over time and with more correctness. However, the correlations seem generally more significant for intensity than for the f0 measurements. Figures 19, 20, and 21 all show that with practice, students generally speak more loudly ($p < 0.001$) for compartment and boundary phrases as well as repair team addresses. We

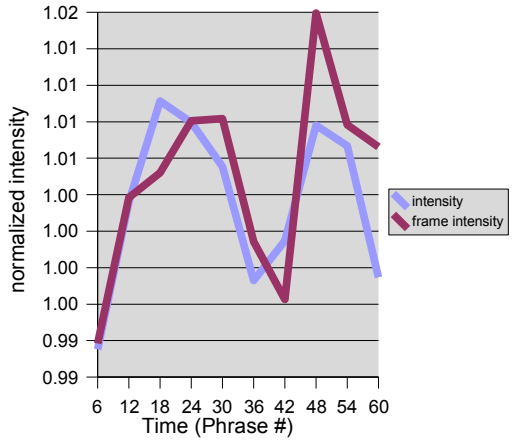

Figure 19: Boundary Phrase Intensity vs. Time

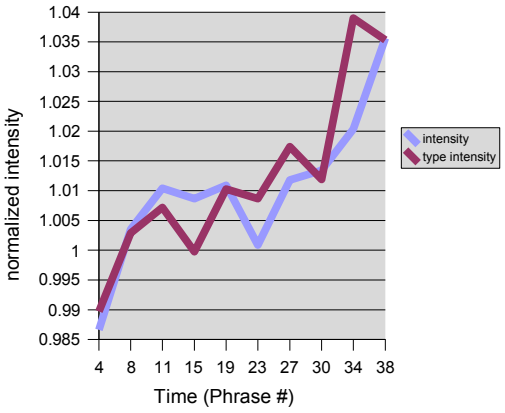

Figure 20: Compartment Phrase Intensity vs. Time

also observe especially significant differences for key words within phrases. Specifically, for boundary frame number, we observe higher intensity for correct phrases (at the $p < 0.001$ level) and increases with practice at the $p < 0.01$ level. For compartment usage type letter, similar correlations with correctness ($p < 0.01$) and practice ($p < 0.001$) are seen. Similar differences are shown for correctness (Figure 22), as correct phrases tend to have greater intensity. This correlation between higher intensity and confidence may point to the use of a sort of "command voice," as students role play issuing orders. Alternatively, it may simply indicate that less confident speech is quieter, particularly surrounding the

items of least confidence.

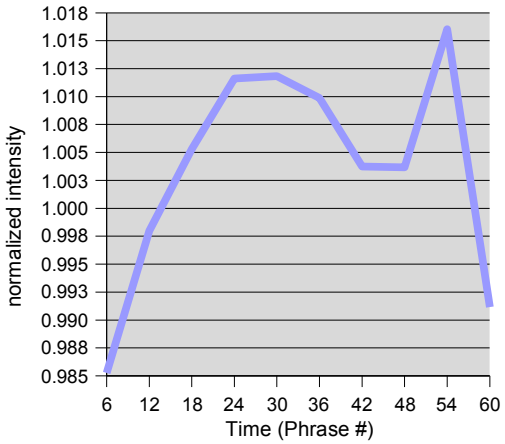

Figure 21: Repair Team Intensity vs. Time

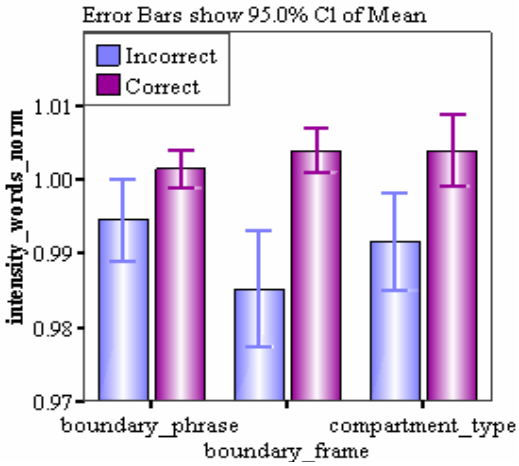

Figure 22: Intensity vs. Correctness

FROM FEATURES TO CLASSIFIER

Naïve Bayesian Overview

The statistical trends relating the various prosodic features of a phrase to its correctness score can be exploited in building a classifier. This section describes the naïve Bayesian classifier employed for this purpose, roughly following the standard formulation for a binary classifier but with some important deviations.

Each phrase can be represented as a vector of numeric values specifying the pitch, intensity, pause rate and so on, and the phrase's associated correctness score can be used as a class label of either *Correct* or *Incorrect*.

$$Y = \langle Y_1, Y_2, \ldots Y_n \rangle$$
$$C \in \{Correct, Incorrect\}$$

The posterior probability of a class given the data can be computed using Bayes' Theorem by multiplying the likelihood by the ratio of priors.

$$P(C|Y) = P(Y|C)\,P(C)/P(Y)$$

Thus, the best class label can be simply defined as the one with the highest probability given the data (i.e., the phrase's feature vector). Furthermore, since there are only two possible classes to consider, a mathematical simplification can be made by computing the

ratio of posteriors and comparing the result with 1.

$$\frac{P(Correct|Y)}{P(Incorrect|Y)} = \frac{P(Y|Correct)P(Correct)}{P(Y|Incorrect)P(Incorrect)}$$

$$classifier(Y) = \left\{ \begin{array}{l} \frac{P(Correct|Y)}{P(Incorrect|Y)} > 1 : Correct, \\ \frac{P(Correct|Y)}{P(Incorrect|Y)} < 1 : Incorrect, \\ otherwise : Indeterminate \end{array} \right\}$$

The third case, that of equal posterior estimates for both classes, rarely occurs in practice. However, an error margin may be chosen so that instead of testing equivalence with 1, a test is made for inclusion within some interval about 1. The size of the interval may be adjusted to increase confidence in the classifications falling outside the indeterminate set, so that the least likely estimates are subsumed in some third set of indeterminate phrases. Note that when defining the interval, one cannot simply employ a linear ε-interval, since the ratio results in a non-linear relationship with the two probabilities, and would result in skewing the indeterminate set toward the class represented in the denominator. However, this can be remedied by either inverting the ratio for values above 1.0 and testing the inverted ratio for inclusion in a linear ε-interval, or, equivalently, all probabilities can be converted to log probabilities.

Even after converting to log probabilities, however, it is still possible that the density of phrases may differ on either side of the center. This argues for a separate ε value for the positive and negative log probability ratios, resulting in the following modified test.

$$classifier(Y) = \left\{ \begin{array}{l} \log\left(\frac{P(Correct|Y)}{P(Incorrect|Y)}\right) > \varepsilon_0 : Correct, \\ \log\left(\frac{P(Correct|Y)}{P(Incorrect|Y)}\right) < -\varepsilon_1 : Incorrect, \\ otherwise : Indeterminate \end{array} \right\}$$

One way of determining appropriate values for the epsilons is to rank the phrases by their probability ratios and then set the epsilons such that it excludes some percentage of the least authoritative judgments. The two epsilons may be set according to separate criteria if the cost of an erroneous judgment is different for the two phrase labels, or alternatively the center can be moved from 1.0. However, for the experiments discussed in this paper the center ratio was left at 1.0 and the epsilons were always adjusted to exclude an equal proportion of the positive and negative judgments.

The prior probability of a given class can be estimated directly from the data by simply counting the number of phrases with a given label divided by the total number of phrases. Also, in the case of discrete valued features, the likelihood of the data given a particular class can be estimated by counting the number of records with a given value and dividing that by the total number of instances of the class.

$$P(C)=\frac{n(C)}{\sum_{c\in\{Correct,\,Incorrect\}} n(c)}$$

$$P(Y|C)=\frac{n(Y\wedge C)}{n(C)}$$

Computing P(Y|C), the joint conditional probability of the features in the Y vector given the class label, is not trivial. However, it can be simply approximated by using the strong independence assumption central to the naïve Bayesian classifier.

$$P(Y|C)=\prod_{i=1}^{n} P(Y_i|C)$$

Note that while it unlikely that independence genuinely holds in the data, a rough approximation can still be made using this assumption. Error is introduced into the output to the extent that the data violates independence. Such violations occur regularly in applications, but the approach is surprisingly robust, and the performance is often still acceptable.

Correctness History and Practice as Additional Features

In addition to the prosodic and speech production features mentioned thus far, the amount of practice and the history of right answers can be employed by the classifier. The amount of practice as measured for the previous statistical analysis can simply be added as another dimension in the feature vector. Correctness itself cannot be employed as a feature, since it is being used as a class label. However, it is reasonable to assume that the history of correct answers for a given phrase type, up to but excluding the correctness of the current phrase, is related to the phrase's own correctness score, if not identical. There is a considerable literature on modeling student mastery by correctness history. However, as a very simple measure, a count of the number of correct phrases occurring within the *n* preceding phrases can be maintained. Some exploration can be employed to determine the optimal size of the window into the correctness history, and in particular it was found through experimentation (discussed in detail later in the “History Tuning” section) that a four phrase window worked best for repair-compartment utterances while a two phrase history worked best for boundary phrases.

This measure is relatively crude, and the reader is referred to (Conati et al, 2002) for just

one example of a better approach. The primary objective here, however, is a basic demonstration of how correctness histories and prosodic information can be combined in complementary fashion. The crude n-phrase window suffices for this purpose but more effective approaches will likely lead to better results than those reported here.

Discretization, PDE, and Expected Distributions

The probability density estimator (PDE) employed for the conditional probability of the data ($P(Y|C)$) that has been described only works for discrete data. However, most of the prosodic features are actually continuous, not discrete at all. Thus, the PDE requires the integration of discretization logic. One method of doing this is to simply split the range of possible values into equal intervals. However, this can lead to overfitting, and may also result in loss of information when the granularity of the interval is too crude to capture the true picture of the data.

Alternatively, the discrete data PDE can be replaced with an estimator based on the assumption that the data should fit some idealized distribution, such as the normal distribution. Then, this idealized distribution can be employed to directly compute a probability of a given data point, given parameters such as the mean value and variance given the class label. This second approach results in less chance for overfitting, but dependence on possibly overly strong assumptions may also lead to a poor approximation of the data.

As a compromise, the data may be discretized in such a way that that the distribution matches some looser but still sufficiently general assumption, guarding both against the overfitting-prone purely data driven approach and the potential for poor approximation due to invalid assumptions.

Casual inspection of the proportion of correct phrases plotted against our prosodic features reveals that many follow a common pattern, with a rise to some peak and then a decline. This pattern can be captured by a simple discretization algorithm even as it approximates the contours of the data. First, sort the phrases by the feature to be discretized. Second, divide the sorted list of phrases into bins of equal numbers of phrases. Determine the bin of maximal concentration of correct phrases. Using this as a maximum, then merge consecutive bins such that their correctness concentration monotonically increases to this maximum and then monotonically declines after it. This algorithm then produces a curve somewhat resembling a normal curve, but with considerable flexibility for variation. Pseudo-code for the algorithm is displayed below.

1 Sort phrases into non-descending order by the given feature.
2 Divide phrases into bins of equal numbers of consecutive phrases.
3 Determine the bin with the maximal concentration of Correct class.
  3.1 Find local optima by identifying all bins whose concentrations are greater than the combined concentrations of all lesser and greater valued bins.
  3.2 Make the local optima of most extreme concentration the global maximum.
4 Proceeding from the first bin to the bin of optimal concentration, merge consecutive bins until a monotonically increasing function of concentration values is formed.
  4.1 Whenever a bin with a smaller concentration is found, merge it with the preceding bin.
  4.2 If the newly formed bin has a smaller concentration than its preceding bin, merge them. Repeat this step until no more merges occur.
  4.3 Proceed to the next bin and repeat from step 4.1 until reaching the optimum.
5 Proceeding from the bin of optimal concentration to the last bin in the sorted list, merge consecutive bins to form a monotonically non-increasing function of concentration. (Symmetric with step 4)

Pseudo-code for Discretization Algorithm

Figures 23 and 24 illustrate the effect of the algorithm on the subject normalized intensity values of boundary phrases. Figure 23 shows that results of the algorithm after step 2, where the phrases have been sorted by intensity and divided into bins of equal numbers of phrases.[7] It appears from this graph that correctness rises sharply to peak somewhere around a normalized value of 1.0, or exactly when intensity reaches the student's mean intensity level, and then slowly declines at higher intensity levels. After running the entire algorithm, shown in Figure 24, it can be seen that the algorithm successfully finds and preserves the rise, peak, and decline while discretizing into only five bins.

7 For this experiment, 20 bins were used for the initial step, resulting in bins of about 125 phrases each for the boundary phrases and 95 phrases each for the repair-compartment utterances.

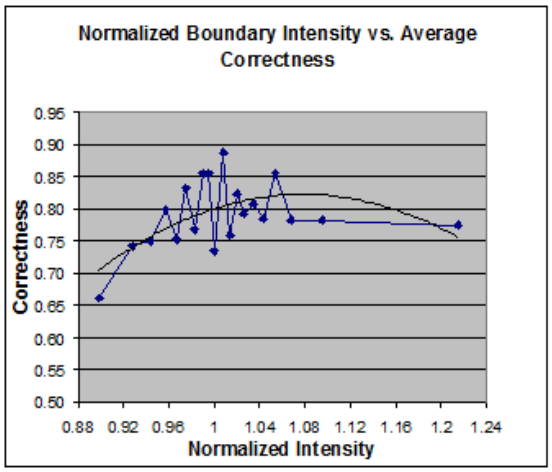

Figure 23: Boundary Intensity vs. Correctness

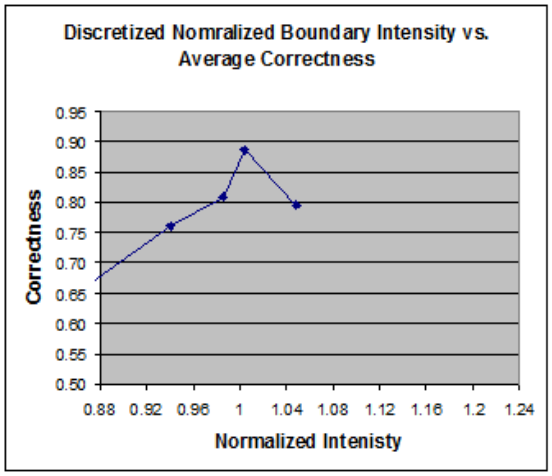

Figure 24: Discretized Boundary Intensity

The algorithm is integral to the probability density estimation step and is essentially a variety of clustering algorithm, which exploits class labels in an attempt to maximize the difference in phrase concentrations between clusters. Making use of the labels is only valid since it is fully integrated into the PDE and trained and tested along with the rest of the classifier. That is, the bins must be defined by value ranges discovered according to the discretization algorithm during training, ranges that can then be used for assigning a data point to its respective bin.

It should be noted that there are many different ways to discretize the data and this section only described one possible approach, a relatively simple approach that was found at least somewhat effective for the task at hand. However, a more thorough investigation should involve a more systematic comparison of different schemes.

Correctness as Confidence Class and the Neutral Set

In the discussion of using correctness for finding relationships with confidence, several issues were noted that may partially obscure the relationship. These issues must, of course, also be acknowledged when employing correctness as confidence class labels. Specifically, noise is introduced by the fact that, with respect to any voluntary uncertainty cues, student confidence is tied to correctness only in as far as the student has an accurate understanding of his own correctness. Noise is further introduced by utterances exhibiting uncertainty regarding non-correctness based issues, such as a student exhibiting uncertainty as to whether the ASR is likely to correctly recognize his current utterance. Furthermore, there may be other emotions that correspond with correctness such as waxing and waning enthusiasm or frustration, which may complicate relationships between prosody and correctness. These noise factors all put an upper bound on the performance of the correctness based classifier.

Nevertheless, certain advantages may potentially outweigh these limitations. In domains where correctness can be automatically determined, significant costs can be saved by forgoing expensive hand generation of confidence labels. Furthermore, such automated methods may be less prone to subjective judgments, as they must be based strictly on machine verifiable standards applied uniformly across all subjects and utterances, potentially mitigating human prejudice. Finally, unlike annotations, machine generated labels can be generated on the fly, allowing the classifier to adapt to new users during a single session, potentially improving classifier performance beyond what is currently possible with hand generated class labels.

Some of these drawbacks may be at least partially overcome by the introduction of a neutrality set, defined by the ε-interval about log probability ratio 0 mentioned previously. By setting the ε-interval, a minimum authority can be specified for classifier output. If a the probability ratio falls within the interval, the classifier instead outputs an indeterminacy flag. Seeing this flag, a tutoring system can refrain from acting on these unlikely guesses. It seems unnecessarily restrictive to force a tutoring system to treat all utterances as either confident or uncertain, as many utterances may be neither, perhaps more accurately characterized as neutral. These utterances would likely fall within the set of utterances of indeterminate classification in the binary classifier, and may be effectively modeled by carefully adjusting the ε values. Thus, the following discussion of classifier performance examines the success rate of the classifier as measured with various settings of the ε threshold.

CLASSIFIER PERFORMANCE

Students generally perform quite well on the two tasks examined. For boundary phrases, students produce the correct information about 78.9% of the time, while for compartment phrases they perform at about 59.5% correctness. These define the prior probabilities of the classifier on the two different phrase types. The task of the classifier is to improve upon these prior probabilities using the likelihood of the data given the class label.

Using 10 fold cross validation testing, the classifier accurately classifies 79.9% of the boundary phrases (barely higher than the prior probability alone) and 69.8% of the repair-compartment phrases. To better evaluate performance it is useful to consider the accuracy of the classifier for each of the different class labels as well as the overall accuracy. Accuracy given that the phrase is correct is commonly referred to as the sensitivity or the true positive rate (tp rate). Accuracy given that the phrase is incorrect is commonly referred to as the specificity or true negative rate (tn rate). Finally, the false positive rate is simply the complement of the tn rate probability, and Receiver Operating Characteristic (ROC) graphs can be used to compare the tn rate and fp rate of different classifiers.

$$tp\ rate = \frac{TP}{TP + FN}$$

$$tn\ rate = \frac{TN}{TN + FP}$$

$$fp\ rate = \frac{FP}{TN + FP} = 1 - tn\ rate$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

Figures 25 an 26 illustrate with ROC graphs the performance of the classifier using three different feature sets: the prosodic and speech production features, the history features, and the combination of these two different feature sets. This type of graph plots the true positive rate versus the false positive rate, where the ideal classifier maximizes the true positive rate while simultaneously minimizing the false positive rate. Thus, the performance of the ideal classifier would appear at the top left corner, where the tp rate is 1.0 and the fp rate is 0.0. The diagonal line portrays the family of random classifiers. For instance, the random classifier that labels the same proportion of correct boundary phrases to incorrect boundary phrases as seen in the data would have true and false positive rates both of 0.789. On the other hand, a classifier that uses only the prior probabilities to decide deterministically would always label phrases as correct, producing true and false positive rates of 1.0, a degenerate random classifier that labels a phrases as "correct" with probability 1.0.

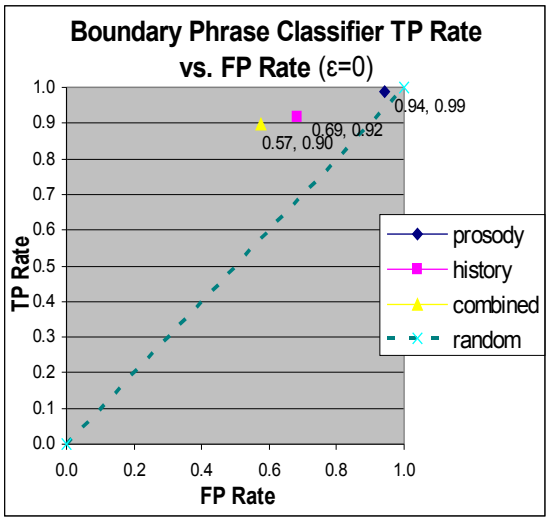

Figure 25: Boundary Phrase Classifier TP Rate vs. FP Rate (ε = 0)

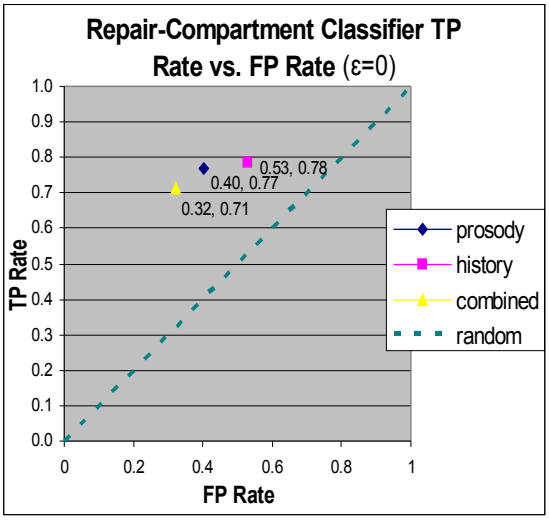

Figure 26: Repair-Compartment Classifier TP Rate vs. FP Rate (ε = 0)

For boundary phrases, the combined classifier has a true positive rate of 0.904 and a false positive rate of 0.521, defining a point well above diagonal. It is clear, then, that the

combined set of features provides a considerable amount of information beyond the basic prior probability of a phrase being correct. All three feature sets classify well above random, distinguishing themselves primarily in how well they classify the incorrect phrases, with history outperforming prosody with a true negative rates of 0.059 vs. 0.315. While prosody alone does a relatively poor job of identifying the incorrect phrases, it outperforms history on the correct phrases with a true positive rate of 0.988 versus the 0.917 of the history based classifier. Figure 26 shows similar performance for repair-compartment utterances, with the principle difference being that the true negative and false positive rates are more closely balanced, though "correct" labels are still slightly favored by all three feature sets. Also, in the case of the repair-compartment utterances prosody actually does a better job of classifying the "incorrect" phrases than history does. What is most important to note, however, is that for both phrase types the combined classifier significantly outperforms both smaller feature sets, demonstrating the utility of prosody as an aid in improving the accuracy of a correctness history only based model.

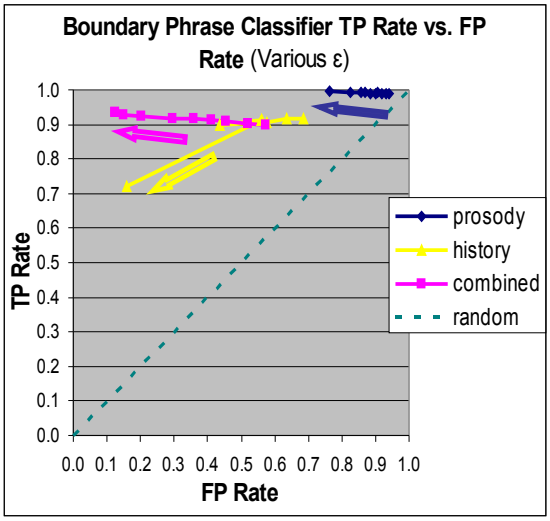

Figure 27: Boundary Phrase Classifier ROC Graph for Various ε Values

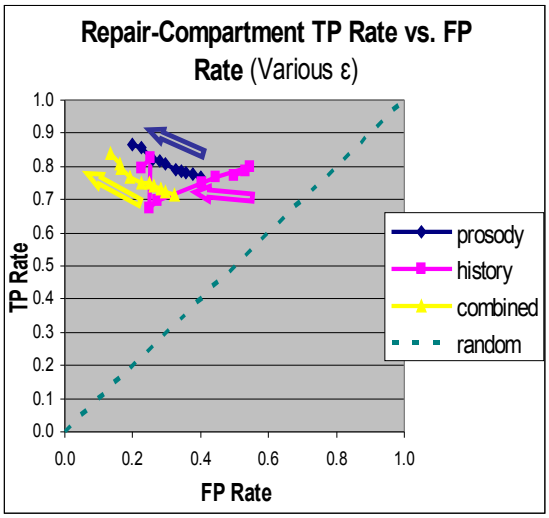

Figure 28: Repair-Compartment Phrase Classifier ROC Graph for Various ε Values

It is also instructive to examine the change in classifier performance given various neutral set sizes. Figures 27 and 28 show ROC graphs with points plotted for epsilon factors at steps of 10% from a neutral set size of 0% up to 90%, where only the most authoritative tenth of the classifications are retained for evaluation. The arrows show the direction of motion as the ε thresholds are gradually increased. As expected, they all gradually improve in performance, as the number of false positives decrease considerably even as the already fairly high true positive rate improves.

Another means of measuring performance is to look at the overall accuracy (or success rate) of the classifier. Figures 29 and 30 illustrate how the accuracy numbers for the

different feature sets improve as more of the classifiers' least authoritative judgments are discarded. As a point of reference, the x-axis is set at the level of the prior probability, since in some sense the classifier should outperform a simplistic classifier that employs a prior only decision rule. Such a classifier would correspond to the situation where the likelihood of the phrase is exactly the same for each class, so that the ratio of conditional probabilities simplifies to the ratio of priors.

$$\frac{P(Correct|Y)}{P(Incorrect|Y)} = \frac{P(Y|Correct)P(Correct)}{P(Y|Incorrect)P(Incorrect)} = \frac{P(Correct)}{P(Incorrect)}$$

That is, the classifier should exceed 78.9% for the boundary phrases, and 59.5% for the repair-compartment utterances. Figures 29 and 30 both show that the classifier does, in fact, improve on these baselines with a 79.9% accuracy for the boundary phrases and a 69.8% accuracy for the repair-compartment utterances. While in the case of the boundary phrases it is only a marginal improvement, the margin widens rapidly as the ε values grow stricter and a larger percentage of the least authoritative judgments are discarded. Thus, when only 30% of the judgments are discarded, the combined boundary classifier achieves an accuracy of 86.0%. After 60% are discarded, accuracy climbs to 90.7%, and so on. The prior of 59.5% for the repair-compartment phrases is easier to exceed, and consequently the gap is considerably larger, however, the smaller prior also lowers overall performance to some extent. Thus, after 30% of the least authoritative judgments are discarded, the classifier achieves a 74.1% accuracy, and after another 30% is discarded it climbs to 78.2%, finally peaking at about 84.9% when all but the top 10% of the most authoritative judgments have been discarded. It is interesting to note that while we observed that the true negative rates are quite different for the prosody and combined feature sets, the overall accuracy is remarkably similar. In fact, for both phrase types,

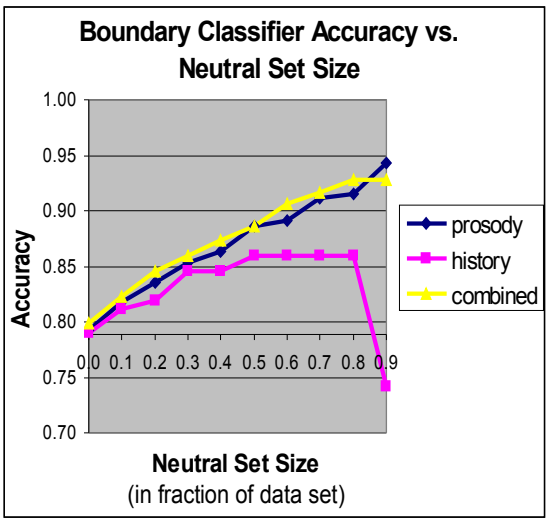

Figure 29: Boundary Classifier Accuracy vs. Neutral Set Size

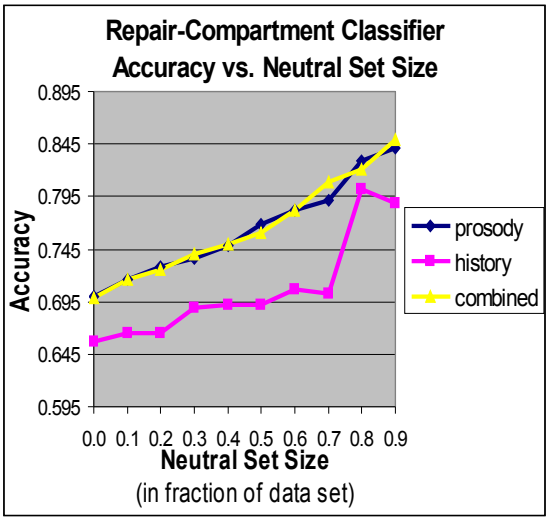

Figure 30: Repair-Compartment Classifier vs. Neutral Set Size

prosody exceeds the history only classifier and is nearly as good as the combined feature set.

While the difference in performance levels differs by about 10 percentage points for the two phrases, the vastly different priors accounts for this difference. To illustrate this point, it is instructive to run the classifier on balanced data sets where the prior probability of a correct phrase is exactly the same as that of an incorrect phrase. This artificial restriction can easily be enforced by randomly removing correct phrases until their number exactly matches that of the incorrect phrases. Figures 31 and 32 illustrate this situation, paralleling figures 29 and 30 for balanced data sets.

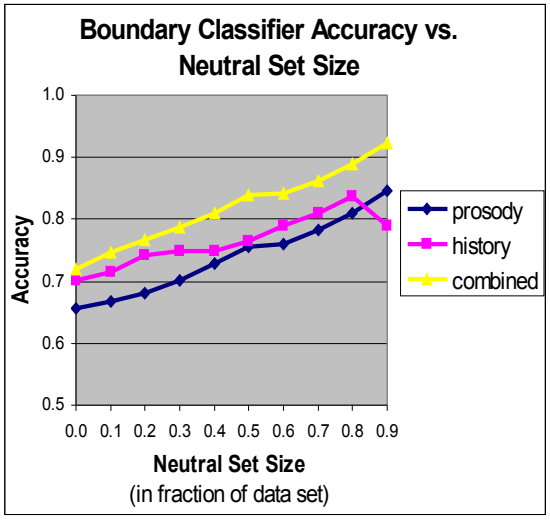

Figure 31: Boundary Classifier vs. Neutral Set Size (Balanced Data Set)

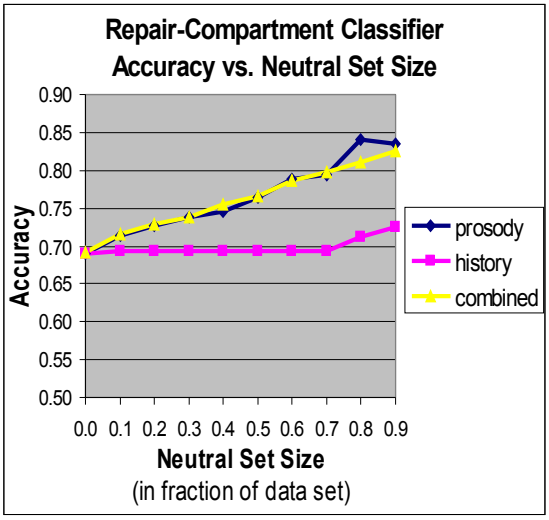

Figure 32: Repair-Compartment Classifier vs. Neutral Set Size (Balanced Data Set)

With balanced data, the two classifiers perform more similarly at an absolute level, both starting at about 70%, though the boundary classifier peaks somewhat higher at 92.4% compared to the 82.6% of the repair-compartment classifier. Aside from illustrating the impact of the prior probabilities, these performance levels further demonstrate the utility of the prosodic feature set, since even in the case where the prior probability offers no information, the prosody only classifier performs at a relatively decent level.

History Tuning

The size of the history window is one area for potential fine tuning. On the one hand, more history means it is possible to observe and make more accurate predictions for students with consistent track records. That is, a student that answered correctly for the last three times is more likely than not to answer the next correctly as well. On the other

hand, a very long history window is more likely to include older information that may no longer be relevant to the current task.[8]

The repair-compartment and boundary phrases effectively demonstrate both principles. For boundary phrases, increases in history length beyond the last two phrases only results in degrading performance. Similarly, so long as the neutral set is empty, classifier performance consistently worsens as the window is lengthened. However, performance increases more rapidly with increasing neutral set size with longer history windows.

The reason for the observed difference in history based classifier performance for boundary phrases and repair-compartment phrases is not obvious, but it seems likely that it is related to the nature of the tasks involved in DCTrain. For repair-compartment phrases, it is very likely that the same repair team and compartment pair will be used in successive commands as the student works through the sequence of tasks required for investigating, isolating, and minimizing the damage in a given compartment. As a result, the student gets several practice opportunities for the same repair team-compartment pair. However, while boundaries do come in sets of four, and the history feature includes some of the boundaries of the same set, each boundary-compartment pair itself generally only occurs once as part of one step among various other quite different tasks. It is possible that this is why history has less bearing on the boundary setting task.

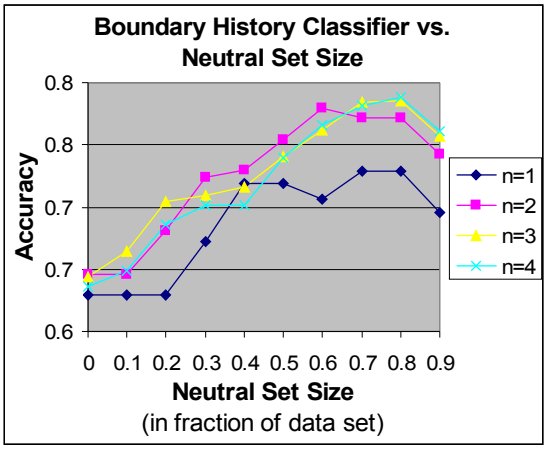

Figure 33: Boundary History Classifier vs. Neutral Set Size

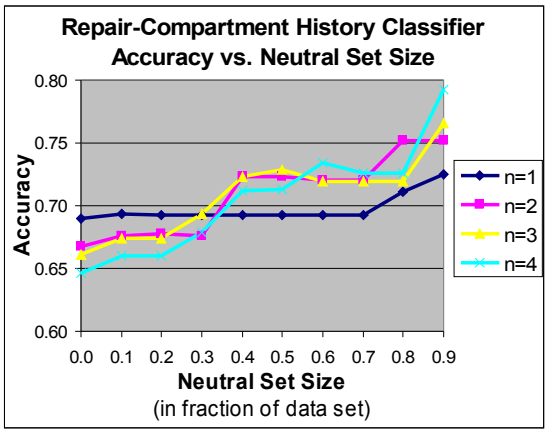

Figure 34: Repair-Compartment History Classifier vs. Neutral Set Size

The ε vs. accuracy graphs for history are observably less continuous than those for the combined or prosody only classifiers. This is an artifact of the discreteness of the probability distribution of the history feature itself. Consider a history window of one, for

8 A different method of tracking correctness history might allow the weighting of older answers less heavily than the more recent answers, allowing a compromise between the more information/relevance of information trade off. This may have been worth exploring given more time.

instance. With information only about whether the previous phrase was correct or incorrect, the probability density estimator can only assume two possible different values. Thus, wherever the ε value is set, it can only break the data at one location, and if the ε value is set lower than that point, the same value is achieved as if it were zero. Similarly, if it is set higher than that point, the effect is the same as it being set at exactly the breaking point. In general, with a binary correctness score, histories of length *n* produce probability functions of 2^n different values with $2^n - 1$ possible breaking points. This effect can be observed from figures 33 and 34 in that shorter histories tend to produce flatter curves with fewer breaks, although it is somewhat obscured by the addition of the practice feature (which results in a more continuous curve).

The natural consequence is that finer grained ε tuning demands longer histories. Hand in hand with this consequence, effective use of very small (or large) ε values also require longer histories. At the same time, one should be aware that a longer history does not always result in improved performance, depending on the nature of the task being modeled.

For all discussions outside this section, the classifier histories were set at two for the boundary phrases and four for repair-compartment phrases.

FEATURE RANKING

In previous sections the effect of prosody and history were examined separately and compared. However, using the balanced data sets and the naïve Bayesian model it is possible to break the model down into its individual constituent features as a means of examining and ranking their individual effects. Figure 35 shows the ranking of the important features for the repair-compartment phrases while Figure 36 shows a similar feature ranking for the boundary phrases. Whereas comparison of means revealed fewer significant features for repair-compartment phrases than for boundary phrases, they proved sufficient for a modestly successful classifier. For both phrase types, correctness history appears the single most effective measure for predicting future student performance. However, it was shown that the combined prosodic features can rival this effectiveness, and even when considered separately they each exhibit better than random performance.

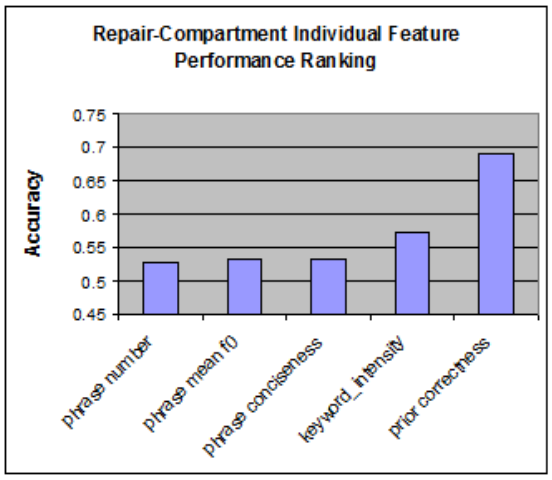

Figure 35: Repair-Compartment Individual Feature Effectiveness

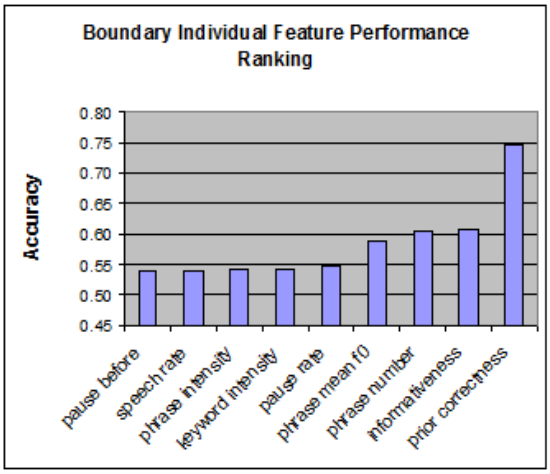

Figure 36: Boundary Individual Feature Effectiveness

CONCLUSIONS AND FUTURE WORK

This work finds both patterns for indicating confidence state and for locating precise items of difficulty within utterances. More pauses before phrases and fewer pauses within phrases, faster speech rate, higher overall intensity and pitch, and falling pitch at the end of phrases all seem to indicate confidence. For precise identification of items of difficulty, these numbers suggest that words articulated with lower intensity, longer durations, and lower f0 are likely to contain the problem. Furthermore, it was observed that pauses are more frequent before portions of phrases that require the most work, perhaps helping to direct attention to the key areas of utterances to analyze with the aid of other features.

Also, utterance length and structure can be very informative, with attention given to how densely the speaker presents information, since concise, information-rich utterances convey the strongest impression of mastery of the material. Furthermore, competent utterances are not only concise and informative but also tend to be organized in logical, clear ways. In the Voice-Enabled DCTrain corpus, for instance, it was observed that students are allowed the flexibility of dividing the four required boundary phrases into as many utterances as they choose. This organizational choice provides more valuable diagnostic information than would be available if they were constrained to present all information items either individually, one per utterance, or all at once. Furthermore, students may include optional information, as in the case of the “net eighty” in repair team addresses, and this provides yet more diagnostic information. Leveraging such information about the student's organizational choices requires analysis of the domain,

and may be facilitated by a careful design of the language interface.

Aside from these findings, one key contribution of this work is the proposal of using automatically extractable measurements of correctness and amount of practice in order to measure confidence. They were employed not only for statistical analysis, but also for bootstrapping machine learning of student confidence, replacing manual annotations for supervised learning and automatic classification.

Further work might include exploring alternative classification algorithms. In particular, the independence assumption of the naïve Bayesian approach very likely degrades performance. While the robustness of the naïve Bayesian approach prevents it from being overwhelmed with error, the statistical analysis nevertheless demonstrated a violation of the independence assumption, particularly between the history and prosodic features. As a result, while the approximation appears sufficiently valid for the modest performance described in this work, dependencies between features suggests relaxing the strong independence assumption. Instead, perhaps a Bayesian net scheme for computing joint probability distributions for multiple simultaneous features would be beneficial.

In addition, the relationship between correctness and student confidence merits closer examination. On the surface level, there is no obvious relationship between correctness and prosody, since correctness itself is not an emotional state, nor a nuance of communication made through intonation and pausing. Considering this, it is intriguing to observe the considerable effectiveness of a classifier based solely on prosody, demonstrating that the correctness of an utterance may often be judged relatively accurately without knowing anything of the content of the utterance.

Furthermore, confidence is only one among many possible affective factors relating to correctness. A more careful factoring of phenomena with a correspondingly more specific prosodic characterization of each would likely yield stronger performance. Careful experiment design and corpus annotation could assist in this work.

Yet another area for potential improvement of the system comes in the treatment of the different phrases as completely different types of data. While the classifiers featured in this paper benefit somewhat from special tailoring to the individual characteristics of the two different student tasks, it may be possible that commonalities and relationships between the two tasks could be exploited for improved classifier performance. For instance, it was observed that the different phrases exhibited somewhat different phenomena relating to correctness and practice, but commonalities were also uncovered in the analysis. A more subtle analysis may uncover a more general explanation that could predict both the similarities and differences. In such a case, it might be possible that a classifier could train on all phrase types simultaneously, obviating the need for separate

classifiers for each, producing a single classifier that can benefit from the larger training set produced by the pooling of the different phrases.

REFERENCES

Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002). Adding human-provided emotional scaffolding to an automated Reading Tutor that listens increases student persistence. Poster presented at Intelligent Tutoring Systems (ITS) Conference, Biarritz, France, June 5-7.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog. In Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado, vol. 3, pp. 2037-2040.

Bennett, C., & Rudnicky, A.I (2002). The Carnegie Mellon Communicator Corpus, In Proceedings of the International Conference of Spoken Language Processing, Denver, Colorado, pp. 341-344. http://www.speech.cs.cmu.edu/sphinx/models/

Berthold, A., & Jameson, A. (1999). Interpreting symptoms of cognitive load in speech input. In Judy Kay, editor, UM99, User Modeling: Proceedings of the Seventh International Conference, pp. 235-244. Springer Wien New York, Vienna.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring, Educational Researcher, no. 13, pp. 4–16.

Boersma, P., & Weeninck, D. (1996). Praat, a system for doing phonetics by computer. Technical Report 132, University of Amsterdam, Inst. of Phonetic Sc.

Bulitko, V. V., & Wilkins, D. C. (1999). Automated instructor assistant for ship damage control. In Proceedings of the Eleventh Conference on Innovative Applications of Artificial Intelligence, pp. 778-785.

Ching, M. (1982). The question intonation in assertions. American Speech, vol. 57, pp. 95–107.

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speech. Cognition, vol. 84, pp. 73-111.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics, In Knowing, learning and instruction:

Essays in honor of Robert Glaser, L. B. Resnick, Ed., pp. 453–494. Lawrence Erlbaum Associates, Hillsdale, NJ.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainly in student modeling. User Modeling & User-Adapted Interaction, vol. 12(4), pp. 371-417.

Core, M. G., Moore, J. D., & Zinn, C. (2003). The role of initiative in tutorial dialogue. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, vol. 1, pp. 67-74.

Evens, M., & Michael, J. (2006). One-on-one Tutoring by Humans and Computers, Lawrence Erlbaum Associates Inc., Mahwah, New Jersey.

Forbes-Riley, K., & Litman, D. (2007). Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development. In Proceedings of Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal, September.

Hirschberg, J., Litman, D., Swerts, M. (2004). Prosodic and other cues to speech recognition failures.- In: Speech communication, vol. 43, pp. 155-175.

Huang, X., Alleva, F., Hwang, M., & Rosenfeld, R. (1993). An overview of the SPHINX-II speech recognition system, In Proceedings of the workshop on Human Language Technology, March 21-24, Princeton, New Jersey, pp. 81-86.

Huang, Z., Chen, L., & Harper, M. (2006). Purdue Prosodic Feature Extraction Toolkit on Praat. Spoken Language Processing Lab, Purdue University, ftp://ftp.ecn.purdue.edu/harper/praat-prosody.tar.gz, March.

Huang, Z., Chen, L., & Harper, M. (2006). An Open Source Prosodic Feature Extraction Tool. In Proceedings of Language Resource and Evaluation Conference (LREC), Genoa, Italy, May.

Litman, D., & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. Companion Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL), Boston, MA, pp. 233-236.

Müller, C., Grossmann-Hutter, B., Jameson, A., Rummer, R. & Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Proceedings of the 8th International Conference on User Modeling (UM2001), pp. 24–33.

Nass, C & Reeves, B. (1996). The Media Equation: How People Treat Computers, Televisions, and New Media as Real People and Places. Cambridge University Press.

Nass, C., & Brave, S. (2005). Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship, MIT Press.

Peters, S., Bratt, E. O., Clark, B., Pon-Barry, H., Schultz, K. (2004). Intelligent Systems for Training Damage Control Assistants, In Proceedings of Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando FL, USA.

Pon-Barry, H., Schultz, K., Bratt, E. O., Clark, B., & Peters, S. (2006). Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems, In International Journal of Artificial Intelligence in Education (IJAIED), vol. 16, 171-194. Special Issue "Best of ITS 2004".

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E. O., & Peters, S. (2004). Advantages of Spoken Language in Dialogue-based Tutoring Systems, In Proceedings of 7th International Conference on Intelligent Tutoring Systems. Maceio, Brazil. Lecture Notes in Computer Science 3220 Springer, ISBN 3-540-22948-5. pp. 390-400.

Rosé, C., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.). Artificial Intelligence in Education, pp. 256-266. Amsterdam: IOS Press.

Rotaru, M., & Litman, D. (2006). Discourse Structure and Speech Recognition Problems. Ninth International Conference on Spoken Language Processing (Interspeech/ICSLP), pp. 53-56.

Shechtman, N., & Horowitz, L. M. (2003). Media inequality in conversation: how people behave differently when interacting with computers and people, Conference on Human Factors in Computing Systems, Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 281–288, Ft. Lauderdale, Florida, USA.

Shinozaki, T., & Furui, S., (2001). Error Analysis Using Decision Trees in Spontaneous Presentation Speech Recognition, In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU2001), Madonna di Campiglio, Trento, Italy, pp. 198-201.

Shinozaki, T., & Furui, S. (2002). Analysis on Individual Differences in Automatic Transcription of Spontaneous Presentations, In Proceedings of the IEEE International

Conference on Acoustics, Speech, and Signal Processing, Proceedings, ICASSP2002, Orlando, U.S.A., vol.1, pp.729-732.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? In M. Swerts and J. Hirschberg (eds.) Special Double Issue on Prosody and Conversation. Language and Speech, vol. 41(3-4), pp. 439-487.

Shriberg, E., & Stolcke, A. (2004). Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing. In Proceedings of the International Conference on Speech Prosody, Nara, Japan, pp. 575-582.

Shute, V.J. (1993). A macroadaptive approach to tutoring, In International Journal of Artificial Intelligence in Education, vol.4, pp. 61-93.

Soltau, H., Waibel, A. (1998), On the influence of hyper-articulated speech on recognition performance, In Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia.

Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to Data Mining, Addison-Wesley Publishing, Boston, Massachusetts.

VanLehn, K. (1988), Student modeling, in Foundations of Intelligent Tutoring Systems, M. Polson & J. Richardson, Eds., pp. 55-78. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wightman, C.W.; Shattuck-Hufnagel, S.; Ostendorf, M.; Price, P.J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. In Journal for the Acoustical Society of America (JASA), 91: pp. 1707-1717.

Zhang, T., Hasegawa-Johnson, M., & Levinson, S. E. (2003). An empathic-tutoring system using spoken language. In Australian Conference on Computer-Human Interaction (OZCHI 2003).