Detecting Spoken Corrections through Decision Tree Methods

Gina-Anne Levow
MIT AI Laboratory
Room 769, 545 Technology Sq
Cambridge, MA 02139
gina@ai.mit.edu

Abstract
Miscommunication in speech recognition systems is unavoidable, but a detailed characterization of user corrections will enable speech systems to identify when a correction is taking place and to more accurately recognize the content of correction utterances. In this paper we exploit the acoustic adaptations made by users during error correction interactions with spoken language systems in the design of decision trees to identify corrections in contrast with original input. Analysis of more than 300 pairs of original and repeat correction utterances revealed significant increases in total utterance and pause durations for all correction types, and increases in pitch variability for corrections of misrecognition errors. We achieve accuracy rates of 75% in distinguishing corrections of misrecognition errors from original inputs. Overall success rates using the trained decision tree classifiers for separating corrections from originals fall between 67-70%. Features based on duration and speaking rate measures proved most useful in the classification. The success at identifying corrections related to misrecognitions is particularly encouraging since misrecognition errors, which produce a recognition result, albeit an incorrect one, are the most difficult for current systems to detect in free-form voice interfaces, often requiring inference.

Content Areas: natural language processing, spoken language understanding, decision-tree learning

Introduction
The frequent recognition errors which plague speech recognition systems present a significant barrier to widespread acceptance of this technology. The difficulty of correcting system misrecognitions is directly correlated with user assessments of system quality. The increased probability of recognition errors immediately after an error compounds this problem. Thus, it becomes crucially important to characterize the differences between original utterances and user corrections of system recognition failures both in order to recognize when a user attempts a correction, indicating a prior recognition error, and to improve recognition accuracy on these problematic utterances. Analysis of data drawn from a field trial of a telephone-based voice-in/voice-out conversational system demonstrates significant differences between original inputs and corrections in measures of duration, pause, and pitch. These differences in turn aid in the development of decision trees which distinguish between new input and user corrections.

Overview
We begin by describing the framework in which the utterances used for these experiments were collected. We then describe the acoustic features used in the data analysis and decision tree design and motivate this choice by considering some related work. A report of the results of the descriptive analysis follows and demonstrates the significant differences among original inputs and different types of corrections. We then turn to the issues involved in the actual construction and assessment of decision trees which identify spoken corrections. The best of these classifiers achieves 75% success rates. Finally, we conclude with a plan to use these results to improve recognition accuracy on spoken corrections and ease recovery from human-computer miscommunication.

System, Subjects, and Errors
The utterances used in the design and analysis of the decision tree classifiers were drawn from approximately 60 hours of user interactions in a field trial of the Sun Microsystems SpeechActs system. (Yankelovich, Levow, & Marx 1995). SpeechActs provides a voice-only interface to common desktop applications such as e-mail, calendar, weather reports, stock quotes, and time and currency conversions. The system allows conversational interactions with speech recognition provided by BBN’s HARK recognizer and speech synthesis through Centigram’s TruVoice system. A sample interaction appears below.

SpeechActs: You have 14 new messages and 47 old messages. The first new message is from Eric Baatz regarding “report completed.”

User: Let me hear it.

SpeechActs: ”The first draft is ready for your comments.
Eric."

User: Reply and include the current message.
SpeechActs: Preparing message to Eric Baatz. Begin recording after the tone. When finished, pause for several seconds.

User: Eric, I'll get those to you this afternoon.
SpeechActs: Say cancel, send, or review.
User: Send.
SpeechActs: Message sent. What now?
User: Next message.
SpeechActs: New message two from Stuart Adams, subject "Can we meet at 3:00 today?"
User: Switch to calendar...

1 The field trial involved a group of nineteen subjects. Four of the participants were members of the system development staff, fourteen were volunteers drawn from Sun Microsystems' sales, marketing, and programming staff, and a final class of subjects consisted of users who called in to experiment with the system. There were three female and sixteen male subjects.

All interactions with the system were recorded and digitized in standard telephone audio quality format at 8kHz sampling in 8-bit mono-law encoding during the conversation. In addition, speech recognition results, parser results, and synthesized responses were logged. A paid assistant then produced a correct verbatim transcript of all user utterances and, by comparing the transcription to the recognition results, labeled each utterance with one of four accuracy codes as described below.

- OK: recognition correct; action correct
- Error Minor: recognition not verbatim, but action correct
- Error: recognition incorrect; action incorrect
- Rejection: no recognition result obtained; no action taken

Overall there were 3096 user utterances recorded, of which 651 resulted in a label of either 'Error' or 'Rejection'; giving an error rate of 21%. 413 utterances, two-thirds of the errors, produced outright rejections, while 227 errors were substitution misrecognitions. The remainder of the errors were due to system crashes or parser errors. The probability of experiencing a recognition failure after a correct recognition was 16%, but immediately after an incorrect recognition it was 44%, 2.75 times greater. This increase in error likelihood suggests a change in speaking style which diverges from the recognizer’s model. The remainder of this paper will identify common acoustic changes which characterize this error correction speaking style. This description leads to the development of a decision tree classifier which can label utterances as corrections or original input.

Related Work
Since full voice-in/voice-out spoken language systems have only recently been developed, little work has been done on error correction dialogs in this context. Two areas of related research that have been investigated are the identification of self-repairs and disfluencies, where the speaker self-interrupts to change an utterance in progress, and some preliminary efforts in the study of corrections in speech input.

In analyzing and identifying self-repairs, (Bear, Dowding, & Shriber 1992) and (Heeman & Allen 1994) found that the most effective methods relied on identifying shared textual regions between the reparandum and the repair. However, these techniques are limited to those instances where a reliable recognition string is available; in general, that is not the case for most speech recognition systems currently available. Alternative approaches described in (Nakatani & Hirschberg 1994) and (Shriber, Bates, & Stolcke 1997), have emphasized acoustic-prosodic cues, including duration, pitch, and amplitude as discriminating features.

The few studies that have focused on spoken corrections of computer misrecognitions, (Oviatt et al. 1996) and (Swerts & Ostendorf 1995), also found significant effects of duration, and in Oviatt et al., pause insertion and lengthening played a role. However, in only one of these studies was input “conversational”, the other was a form-filling application, and neither involved spoken system responses, relying instead on visual displays for feedback, with potential impact on speaking style.

Error Data, Features, and Examples
For these experiments, we selected pairs of utterances: the first (original) utterance is the first attempt by the user to enter an input or a query; the second (repeat) follows a system recognition error, either misrecognition or rejection, and tries to correct the mistake in the same words as the original. For example:

SYSTEM SAID: Please say mail, calendar, weather, stock quotes or start over to begin again.

SYSTEM HEARD: MAIL
USER SAID: MAIL
CODE: OK
SYSTEM SAID: Switching to mail.
SYSTEM SAID: You have 495 old messages.
SYSTEM SAID: and 8 urgent messages.
SYSTEM SAID: and one new message.
SYSTEM SAID: Your first message is from
Solaris TOI Admin regarding
SOLARIS 2.5 REGISTRATION.

SYSTEM HEARD: “nothing”
USER SAID: Read Message Four Eight Nine
CODE: Rejection

SYSTEM SAID: Sorry?
SYSTEM HEARD: “nothing”
USER SAID: Read Message Four Eight Nine
CODE: Rejection

SYSTEM SAID: Still no luck. Speak clearly, but don’t over emphasize.

SYSTEM HEARD: Go To Message Four Please Umm
USER SAID: Go To Message Four Eight Nine
CODE: ERROR

SYSTEM SAID: I heard goto new message 4. You only have 1 new
messages and you are currently on new message 1.

SYSTEM HEARD: Go To Old Message
Forty Nine
USER SAID: Go To Old Message
Four Eight Nine
CODE: ERROR

SYSTEM SAID: Skipping to old message 49.
SYSTEM SAID: Old message 49 From Randy Lewis.
SYSTEM SAID: Subject Q2 Commissions Information.

In total, there were 302 of these original-repeat pairs: 196 resulting from rejections, and 106 from misrecognitions.

Following (Oviatt et al. 1996), (Shriberg, Bates, & Stolcke 1997), and (Ostendorf et al. 1996), we coded a set of acoustic-prosodic features to describe the utterances. These features fall into four main groups: duraional, pause, pitch, and amplitude.

Duration

The basic duration measure is total utterance duration. This value is obtained through a two-step procedure. First we perform an automatic forced alignment of the utterance to the verbatim transcription text using the OGI CSLU CSLUst Toolkit (Colton 1995). Then the alignment is inspected and, if necessary, adjusted by hand to correct for any errors, such as those caused by extraneous background noise or non-speech sounds. A typical alignment appears in Figure 1. In addition to the simple measure of total duration in milliseconds, a number of derived measures also prove useful. Some examples of such measures are speaking rate in terms of syllables per second and a ratio of the actual utterance duration to the mean duration for that type of utterance.

Pause

A pause is any region of silence internal to an utterance and longer than 10 milliseconds in duration. Silences preceding unvoiced stops and affricates were not coded as pauses due to the difficulty of identifying the onset of consonants of these classes. Pause-based features include number of pauses, average pause duration, total pause duration, and silence as a percentage of total utterance duration. An example of pause insertion and lengthening appear in Figure 1.

Pitch

To derive pitch features, we first apply the $F_0$ (fundamental frequency) analysis function from the Entropic ESPS Waves+ system (Secrest & Doddington) to produce a basic pitch track. A trained analyst examines this track to remove any points of doubling or halving due to pitch tracker error, non-speech sounds, and excessive glottalization of $\geq 5$ sample points. We compute several derived measures using simple algorithms to obtain $F_0$ maximum, $F_0$ minimum, $F_0$ range, final $F_0$ contour, slope of maximum pitch rise, slope of maximum pitch fall, and sum of the slopes of the steepest rise and fall. Figure 1 depicts a basic pitch contour.

Amplitude

Amplitude, measuring the loudness of an utterance, is also computed using the ESPS Waves+ system. Mean amplitudes are computed over all voiced regions with amplitude $\geq 30 \text{dB}$. Amplitude features include utterance mean amplitude, mean amplitude of last voiced region, amplitude of loudest region, standard deviation, and difference from mean to last and maximum to last.

Descriptive Acoustic Analysis

Using the features described above, we performed some initial simple statistical analyses to identify those features which would be most useful in distinguishing original inputs from repeat corrections, and corrections of rejection errors (CRE) from corrections of misrecognition errors (CME). The results for the most interesting features, duration, pause, and pitch, are described below.

Duration

Total utterance duration is significantly greater for corrections than for original inputs. In addition, increases in correction duration relative to mean duration for the utterance prove significantly greater for CME's than for CRE's.
Figure 1: Above: A lexically matched pair where the repeat (bottom) has an 18% increase in total duration and a 400% increase in pause duration.
Below: Contrasting Falling (top) and Rising (bottom) Pitch Contours
Pause
Similarly to utterance duration, total pause length increases from original to repeat. For original-repeat pairs where at least one pause appears, paired t-test on log-transformed data reveal significantly greater pause durations for corrections than for original inputs.

Pitch
While no overall trends reached significance for pitch measures, CRE's and CME's, when considered separately, did reveal some interesting contrasts between corrections and original inputs within each subset and between the two types of corrections. Specifically, male speakers showed a small but significant decrease in pitch minimum for CRE's.

CME's produced two unexpected results. First they displayed a large and significant increase in pitch variability from original to repeat as measured by the sum of the slopes of the steepest rise and fall. In addition, they also showed significant increases in steepest rise, steepest fall, and sum of slopes measures when compared with CRE's.

Decision Tree Experiments
The next step was to develop predictive classifiers of original vs repeat corrections and CME's vs CRE's informed by the descriptive analysis above. We chose to implement these classifiers with decision trees (using Quinlan's (Quinlan 1992) C4.5) trained on a subset of the original-repeat pair data. Decision trees have two features which make them desirable for this task. First, since they can ignore irrelevant attributes, they will not be misled by meaningless noise in one or more of the 38 duration, pause, pitch, and amplitude features coded. Since these features are probably not all important, it is desirable to use a technique which can identify those which are most relevant. Second, decision trees are highly intelligible; simple inspection of trees can identify which rules use which attributes to arrive at a classification, unlike more opaque machine learning techniques such as neural nets.

Feature Modification: Absolute to Relative
A significant hurdle still remains in adapting the descriptive features used in the statistical analysis to features suitable for building a prescriptive classifier. The problem we encounter is that, although there are significant increases in the means of duration, pause, and pitch measures, the individual values of these attributes vary even more. For example, total utterance duration on average increases 10% from original to repeat correction, but since utterances range from 100 to 3340 milliseconds in duration due to different lexical content, there will be some original utterances longer than most repeats. To counter this difficulty, we adopted measures that used relative rather than absolute measures. In other words instead of using an absolute measure such as utterance or pause duration or amplitude, we used measures which emphasize relative changes in length from a mean utterance length or relative changes in amplitude from a mean speaker amplitude. Modifications for duration and amplitude measures proved simplest, resulting in an array of features such as speaking rate in syllables per second and a ratio of utterance length to mean length for that utterance. Modification of pitch-related features has proven more difficult. This situation arises in part because speaker and sex-based differences in baseline F0 and pitch range do not allow for simple linear transformations like those applied to amplitude. This issue presents a challenge for future research.

Decision Trees: Results & Discussion
The first set of decision tree trials attempted to classify original and repeat correction utterances, for both correction types. We used a set of 38 attributes: 18 based on duration and pause measures, 6 on amplitude, five on pitch height and range, and 13 on pitch contour. Trials were made with each of the possible subsets of these four feature classes on over 600 instances with seven-way cross-validation. The best results, 33% error, were obtained using attributes from all sets. Duration measures were most important, providing an improvement of at least 10% in accuracy over all trees without duration features.

The next set of trials dealt with the two error correction classes separately. One focused on distinguishing CME's from CRE's, while the other concentrated on differentiating CME's alone from original inputs. The test attributes and trial structure were the same as above. The best error rate for the CME vs. CRE classifier was 30.7%, again achieved with attributes from all classes, but depending most heavily on durational features. Finally the most successful decision trees were those separating original inputs from CME's. These trees obtained an accuracy rate of 73% (25% error) using similar attributes to the previous trials. The most important splits were based on pitch slope and durational features. An exemplar of this type of decision tree is shown below.

\[
\text{lexincvowel} > 0.2335 : r (39.0/4.9) \\
\text{lexincvowel} \leq 0.2335 : \\
\quad \text{lenexp} \leq 1.0116 : \\
\quad \quad \text{lexincvowel} > -0.0023 : o (51.0/2.6) \\
\quad \quad \text{lexincvowel} \leq -0.0023 : \\
\quad \quad \quad \text{pitchslope} > 0.265 : o (19.0/3.7) \\
\quad \quad \quad \text{pitchslope} \leq 0.265 : \\
\quad \quad \quad \quad \text{pitchlastmin} \leq 25.2214: r(11/2) \\
\quad \quad \quad \quad \text{pitchlastmin} > 25.2214: \\
\quad \quad \quad \quad \quad \text{minslope} \leq -0.221: r(18/5) \\
\quad \quad \quad \quad \quad \text{minslope} > -0.221: o(18/5) \\
\quad \quad \text{lenexp} > 1.0116 : \\
\quad \quad \quad \maxprop > 0.0615 : r (7.0/1.3) \\
\quad \quad \quad \maxprop \leq 0.0615 : \\
\quad \quad \quad \quad \text{lenexp} \leq 1.0277 : r (8.0/3.5)
\]
\[ \text{lenexp} > 1.0277 : o (19.0/8.0) \]
\[ \text{lexincverpsyll} > 20.471 : \]
\[ \text{pitchslope} \leq 0.281 : r (24.0/3.7) \]
\[ \text{pitchslope} > 0.281 : o (7.0/2.4) \]

These decision tree results in conjunction with the earlier descriptive analysis provide evidence of strong contrasts between original inputs and repeat corrections, as well as between the two classes of corrections. They suggest that different error rates after correct and after erroneous recognitions are due to a change in speaking style that we have begun to model.

In addition, the results on corrections of misrecognition errors are particularly encouraging. In current systems, all recognition results are treated as new input unless a rejection occurs. User corrections of system misrecognition errors can currently only be identified by attempting to reason about whether the current utterance violates some existing belief or presupposition of the system. This reasoning process can be quite complex and depends on the current utterance being correctly transcribed by the system. In contrast, the method described here provides a way to use acoustic features such as duration, pause, and pitch variability to identify these particularly challenging error corrections without strict dependence on a perfect textual transcription of the input and with relatively little computational effort.

Conclusions & Future Work

Using acoustic-prosodic features such as duration, pause, and pitch variability to identify error corrections in spoken dialog systems shows promise for resolving this knotty problem. We further plan to explore the use of more accurate characterization of the contrasts between original and correction inputs to adapt standard recognition procedures to improve recognition accuracy in error correction interactions. Helping to identify and successfully recognize spoken corrections will improve the ease of recovering from human-computer miscommunication and will lower this hurdle to widespread acceptance of spoken language systems.

References


