Characterizing and Recognizing Spoken Corrections in Human-Computer Dialogue

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 investigate the adaptations of users when they encounter recognition errors in interactions with a voice-in/voice-out spoken language system. In analyzing more than 300 pairs of original and repeat correction utterances, matched on speaker and lexical content, we found overall increases in both utterance and pause duration from original to correction. Interestingly, corrections of misrecognition errors (CME) exhibited significantly heightened pitch variability, while corrections of rejection errors (CRE) showed only a small but significant decrease in pitch minimum. CME's demonstrated much greater increases in measures of duration and pitch variability than CRE's. These contrasts allow the development of decision trees which distinguish CME's from CRE's and from original inputs at 70-75% accuracy based on duration, pitch, and amplitude features.

1 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.

2 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.

3 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 et al., 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 current message.
SpeechActs:	Preparing message to Eric Baatz. Begin recording after the tone. When finished, pause 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' staff, and a final class of subjects consisted of one-time guest users 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 mu-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 exact; action correct
Error: recognition incorrect; action incorrect
Rejection: no recognition result; no action

Overall there were 7752 user utterances recorded, of which 1961 resulted in a label of either ’Error’ or ’Rejection’, giving an error rate of 25%. 1250 utterances, almost two-thirds of the errors, produced outright rejections, while 706 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.

4 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 et al., 1992) and (Heeman and 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 and Hirschberg, 1994) and (Shriberg et al., 1997), have emphasized acoustic-prosodic cues, including duration, pitch, and amplitude as discriminating features.

The few studies that have focussed on spoken corrections of computer misrecognitions, (Oviatt et al., 1996) and (Swerts and 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.

5 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.
USER SAID:MAIL
SYSTEM HEARD: MAIL
CODE: OK
SYSTEM SAID: Switching to mail.
Your first message is ..
USER SAID:Read message four eight nine
SYSTEM HEARD: "nothing"
CODE: Rejection
SYSTEM SAID: Sorry ?
USER SAID:Read message four eight nine
SYSTEM HEARD: "nothing"
CODE: Rejection
SYSTEM SAID: Still no luck. Speak clearly,
but don't over emphasize .
USER SAID:Go to message four eight nine
SYSTEM HEARD: Go to message four
please umm
CODE: ERROR
SYSTEM SAID: I heard goto new message 4.

In total, there were 302 of these original-repeat pairs: 214 resulting from rejections, and 88 from misrecognitions.

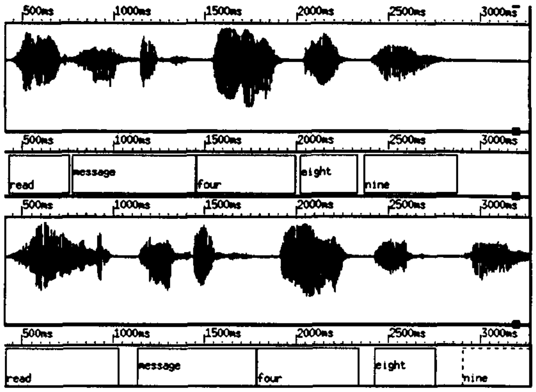

Figure 1: A lexically matched pair where the repeat (bottom) has an 18% increase in total duration and a 400% increase in pause duration.

Following (Oviatt et al., 1996), (Shriberg et al., 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: durational, pause, pitch, and amplitude. We further selected variants of these feature classes that could be scored automatically, or at least mostly automatically with some minor hand-adjustment. We hoped that these features would be available during the recognition process so that ultimately the original-repeat correction contrasts would be identified automatically.

5.1 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 CSLUsh 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.

5.2 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

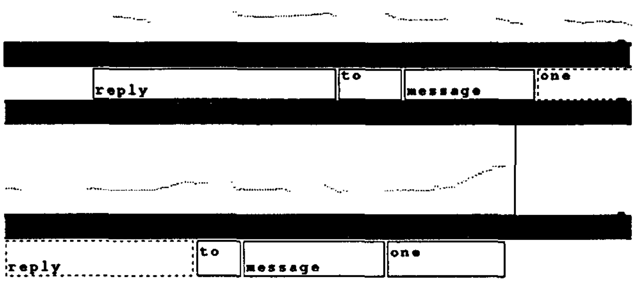

Figure 2: Contrasting Falling (top) and Rising (bottom) Pitch Contours

insertion and lengthening appear in Figure 1.

5.3 Pitch

To derive pitch features, we first apply the F0 (fundamental frequency) analysis function from the Entropic ESPS Waves+ system (Secrest and Doddington, 1993) to produce a basic pitch track. Most of the related work reported above had found relationships between the magnitude of pitch features and discourse function rather than presence of accent type, used more heavily by (Pierrehumbert and Hirschberg, 1990), (Hirschberg and Litman, 1993). Thus, we chose to concentrate on pitch features of the former type. A trained analyst examines the pitch track to remove any points of doubling or halving due to pitch tracker error, non-speech sounds, and excessive glottalization of ≥ 5 sample points. We compute several derived measures using simple algorithms to obtain F0 maximum, F0 minimum, F0 range, final F0 contour, slope of maximum pitch rise, slope of maximum pitch fall, and sum of the slopes of the steepest rise and fall. Figure 2 depicts a basic pitch contour.

5.4 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 $\geq$ 30dB. 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.

6 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.

6.1 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.

6.2 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.

6.3 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 the slope of the steepest rise, while CRE's exhibited a corresponding decrease rising slopes. In addition, they also showed significant increases in steepest rise measures when compared with CRE's.

7 Discussion

The acoustic-prosodic measures we have examined indicate substantial differences not only between original inputs and repeat corrections, but also between the two correction classes, those in response to rejections and those in response to misrecognitions. Let us consider the relation of these results to those of related work and produce a more clear overall picture of spoken correction behavior in human-computer dialogue.

7.1 Duration and Pause: Conversational to Clear Speech

Durational measures, particularly increases in duration, appear as a common phenomenon among several analyses of speaking style [(Oviatt et al., 1996), (Ostendorf et al., 1996), (Shriberg et al., 1997)]. Similarly, increases in number and duration of silence regions are associated with disfluencies (Shriberg et al., 1997), self-repairs (Nakatani and Hirschberg, 1994), and more careful speech (Ostendorf et al., 1996) as well as with spoken corrections (Oviatt et al., 1996). These changes in our correction data fit smoothly into an analysis of error corrections as invoking shifts from conversational to more "clear" or "careful" speaking styles. Thus, we observe a parallel between the changes in duration and pause from original to repeat correction, described as conversational to clear in (Oviatt et al., 1996), and from casual conversation to carefully read speech in (Ostendorf et al., 1996).

7.2 Pitch

Pitch, on the other hand, does not fit smoothly into this picture of corrections taking on clear speech characteristics similar to those found in carefully read speech. First of all, (Ostendorf et al., 1996) did not find any pitch measures to be useful in distinguishing speaking mode on the continuum from a rapid conversational style to a carefully read style. Second, pitch features seem to play little role in corrections of rejections. Only a small decrease in pitch minimum was found, and this difference can easily be explained by the combination of two simple trends. First, there was a decrease in the number of final rising contours, and second, there were increases in utterance length, that, even under constant rates of declination, will yield lower pitch minima. Third, this feature produces a divergence in behavior of CME's from CRE's.

While CRE's exhibited only the change in pitch minimum described above, corrections of misrecognition errors displayed some dramatic changes in pitch behavior. Since we observed that simple measures of pitch maximum, minimum, and range failed to capture even the basic contrast of rising versus falling contour, we extended our feature set with measures of slope of rise and slope of fall. These measures may be viewed both as an attempt to create a simplified form of Taylor's rise-fall-continuation model (Taylor, 1995) and as an attempt to provide quantitative measures of pitch accent. Measures of pitch accent and contour had shown some utility in identifying certain discourse relations [(Pierrehumbert and Hirschberg, 1990), (Hirschberg and Litman, 1993). Although changes in pitch maxima and minima were not significant in themselves, the increases in rise slopes for CME's in contrast to flattening of rise slopes in CRE's combined to form a highly significant measure. While not defining a specific overall contour as in (Taylor, 1995), this trend clearly indicates increased pitch accentuation. Future work will seek to describe not only the magnitude, but also the form of these pitch accents and their relation to those outlined in (Pierrehumbert and Hirschberg, 1990).

7.3 Summary

It is clear that many of the adaptations associated with error corrections can be attributed to a general shift from conversational to clear speech articulation. However, while this model may adequately describe corrections of rejection errors, corrections of misrecognition errors obviously incorporate additional pitch accent features to indicate their discourse function. These contrasts will be shown to ease the identification of these utterances as corrections and to highlight their contrastive intent.

8 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.

8.1 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 focussed 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 75% (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 in shown below.

```
normduration1 > 0.2335 : r (39.0/4.9)
normduration1 <= 0.2335 :
|normduration2 <= 20.471 :
||normduration3 <= 1.0116 :
|||normduration1 > -0.0023 : o (51/3)
|||normduration1 <= -0.0023 :
|||| pitchslope > 0.265 : o (19/4))
|||| pitchslope <= 0.265 :
||||| pitchlastmin <= 25.2214:r(11/2)
||||| pitchlastmin > 25.2214:
|||||| minslope <= -0.221:r(18/5)
|||||| minslope > -0.221:o(15/5)
||normduration3 > 1.0116 :
|||normduration4 > 0.0615 : r (7.0/1.3)
|||normduration4 <= 0.0615 :
||||normduration3 <= 1.0277 : r (8.0/3.5)
||||normduration3 > 1.0277 : o (19.0/8.0)
|normduration2 > 20.471 :
|| pitchslope <= 0.281 : r (24.0/3.7)
|| 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 misrecognitions can currently only be identified by complex reasoning requiring an accurate transcription. 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.

9 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

J. Bear, J. Dowding, and E. Shriberg. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the ACL*, pages 56–63, University of Delaware, Newark, DE.

D. Colton. 1995. Course manual for CSE 553 speech recognition laboratory. Technical Report CSLU-007-95, Center for Spoken Language Understanding, Oregon Graduate Institute, July.

P.A. Heeman and J. Allen. 1994. Detecting and correcting speech repairs. In *Proceedings of the ACL*, pages 295–302, New Mexico State University, Las Cruces, NM.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational linguistics*, 19(3):501–530.

C.H. Nakatani and J. Hirschberg. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustic Society of America*, 95(3):1603–1616.

M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shribergand D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. 1996. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proceedings of the International Conference on Spoken Language Processing.* supplementary paper.

S.L. Oviatt, G. Levow, M. MacEarchern, and K. Kuhn. 1996. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 801–804.

Janet Pierrehumbert and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 271–312. MIT Press, Cambridge, MA.

J.R. Quinlan. 1992. *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

B. G. Secrest and G. R. Doddington. 1993. An integrated pitch tracking algorithm for speech systems. In *ICASSP 1993.*

E. Shriberg, R. Bates, and A. Stolcke. 1997. A prosody-only decision-tree model for disfluency detection. In *Eurospeech '97.*

M. Swerts and M. Ostendorf. 1995. Discourse prosody in human-machine interactions. In *Proceedings of the ECSA Tutorial and Research Workshop on Spoken Dialog Systems - Theories and Applications.*

Paul Taylor. 1995. The rise/fall/continuation model of intonation. *Speech Communication*, 15:169–186.

N. Yankelovich, G. Levow, and M. Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *CHI '95 Conference on Human Factors in Computing Systems*, Denver, CO, May.