CORRECTIONS IN SPOKEN DIALOGUE SYSTEMS

Marc Swerts[1], *Diane Litman*[2] *and Julia Hirschberg*[2]

[1]IPO, Eindhoven, The Netherlands, and
CNTS, Antwerp, Belgium
[2]AT&T Labs–Research, Florham Park, NJ, USA
m.g.j.swerts@tue.nl, {diane/julia}@research.att.com

ABSTRACT

This study analyzes user corrections of system errors in the TOOT spoken dialogue system. We find that corrections differ from non-corrections prosodically, in ways consistent with hyperarticulated speech, although many corrections are *not* hyperarticulated. Yet both are misrecognized more frequently than non-corrections — though no more likely to be rejected by the system. Corrections more distant from the error they correct tend to exhibit greater prosodic differences, and also to be recognized more poorly. System dialogue strategy affects users' choice of correction type, suggesting that strategy-specific methods of detecting or coaching users on corrections may be useful. Strategies that produce longer tasks but fewer misrecognitions and subsequent corrections are preferred by users.

1. INTRODUCTION

Since spoken dialogue systems often make mistakes in recognizing user input, accurate methods of detecting and correcting system errors are essential to supporting successful interactions. Understanding how users attempt to correct system failures and why their attempts succeed or fail is important to improve the design of future spoken dialogue systems. For example, knowing whether they are more likely to repeat or rephrase their utterances, add new information or shorten their input, and how system behavior influences these choices, can suggest appropriate on-line modifications to the system's interaction strategy or to the recognition procedure it employs. Determining which speaker behaviors are more successful in correcting system errors can also lead to improvements in the help information such systems provide.

We are conducting studies focussing on the role prosody may play in both detecting automatic speech recognition (ASR) errors and in helping to understand user corrections of such errors. In two different corpora of human-machine interactions, we found that prosodic features can be used to *detect* recognition errors with considerable accuracy[2, 8, 3]: in combination with information already available to the recognizer, such as acoustic confidence scores, grammar and recognized string, they can distinguish speaker turns that are misrecognized far better than traditional methods for ASR rejection using acoustic confidence scores alone. In the current study, we turn to the question of how people try to *correct* ASR errors in their interactions with machines, and the role that prosody may play in identifying user corrections and in helping to analyze them.

Previous research has shown that users often have difficulty dealing with errors made by current dialogue systems — and that systems also find it hard to handle user attempts to correct them. So, repair strategies in human-machine interactions can be quite ineffective. There is evidence that dialogue confirmation strategies may hinder users' ability to correct system error. For instance, if a system wrongly presents information as being correct, as when it verifies information implicitly, users become confused about how to respond [5]. Other studies have shown that speakers tend to switch to a prosodically 'marked' speaking style after communication errors, comparing repetition corrections with the speech being repeated [13, 11, 7, 1]. While this speaking style may be effective in problematic human-human communicative settings, there is evidence that suggests it leads to further errors in human-machine interactions [7, 12], perhaps because it differs from the speech data most recognizers are trained on.

In this paper, we describe an analysis of user corrections of system error collected in the TOOT spoken dialogue system. In Section 2, we describe the corpus itself and how it was collected and labeled. In Section 3, we characterize the nature of corrections in this corpus, in terms of when they occur, how well they are handled by the system, what distinguishes their prosody from other utterances, their relationship to the utterances they correct, and how they differ according to dialogue strategy. We conclude with some implications of our results for future spoken dialogue systems and some goals of our future research.

2. THE TOOT CORPUS

Our corpus consists of dialogues between human subjects and TOOT, a spoken dialogue system that allows access to train information from the web via telephone. TOOT was collected to study variations in dialogue strategy and in user-adapted interaction [9]. It is implemented using an IVR (interactive voice response) platform developed at AT&T, combining ASR and text-to-speech with a phone interface [4]. The system's speech recognizer is a speaker-independent hidden Markov model system with context-dependent phone models for telephone speech and constrained grammars defining vocabulary at any dialogue state. The platform supports barge-in. Subjects performed four tasks with one of several versions of the system that differed in terms of locus of initiative (system, user, or mixed), confirmation strategy (explicit, implicit, or none), and whether these conditions could be changed by the user during the task (adaptive vs. non-adaptive). Subjects were 39 students; 20 native speakers and 19 non-native, 16 female and 23 male. Dialogues were recorded and system and user behavior logged automatically. The *concept accuracy* (CA) of each turn was manually labeled. If the ASR correctly captured all task-related information in the turn (e.g. time, departure and arrival cities), the turn's CA score was 1 (*semantically correct*). Otherwise, the CA score reflected the percentage of correctly recognized task information in the turn. The dialogues were also transcribed and automatically scored in comparison to the ASR recognized string to produce a *word error rate* (WER) for each turn. For the study described below, we examined 2328 user turns (all user input between two system inputs) from 152 dialogues.

3. TOOT CORPUS CORRECTIONS

To identify corrections in the corpus two authors independently labeled each turn as to whether or not it constituted a correction of a prior system failure (a rejection or CA error, which were the only system failure subjects were aware of) and subsequently decided upon a consensus label. Each turn labeled 'correction' was further classified as belonging to one of the following categories: REP (repetition, including repetitions with differences in pronunciation or fluency), PAR (paraphrase); ADD (task-relevant content added); OMIT (content omitted); and ADD/OMIT (content both added and omitted). For user turns containing a correction plus one or more additional dialogue acts, only the correction is used for purposes of analysis below. We also labeled as *restarts* user corrections which followed non-initial system-initial prompts (e.g. "How may I help you?" or "What city do you want to go to?"); in such cases system and user essentially started the dialogue over from the beginning. Each correction was also indexed with an identifier representing the prior turn it was correcting, so that we could investigate "chains" of corrections of a single failed turn, by tracing back through subsequent corrections of that turn. Figure 1 shows a fragment of a TOOT dialogue with corrections labeled.

Figure 1: Toot Dialogue Fragment with Correction Labels

	Turn	Turn ID	Corr of	Corr Type
S:	*On which day of the week do you want to leave?*			
S:	*Sorry, I didn't hear anything from you. Please tell me your departure date.*			
U:	Um tomorrow	776	NA	NA
S:	*You can say the departure date. For example, you can say 'today' or 'tomorrow'. Or you can say 'Monday', 'Tuesday', 'Wednesday'...*			
U:	Tomorrow	777	776	REP
S:	*You can say the departure date. For example, you can say 'today' or 'tomorrow'. Or you can say 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday'...*			
U:	Saturday	778	777	PAR
S:	*Do you want to leave on Saturday?*			
U:	Yes	779	NA	NA
S:	*What time do you want to leave?*			
U:	Eight o'clock A M	780	NA	NA
S:	*Do you want to leave around ten oclock p-m*			
U:	Eight o'clock	781	780	OMIT
S:	*Do you want to leave around ten oclock p-m*			
U:	No	782	NA	NA
S:	*What time do you want to leave?*			
U:	Eight o'clock A M	783	781	ADD

We identified 29% of the turns in the corpus as corrections. These were a serious problem for the system, being misrecognized (in terms of WER) about twice as often (61%) as non-corrections (33%). Despite this higher misrecognition rate, the ASR system rejected corrections at about the same rate (9% of the time vs. 8%) as non-corrections. So, the system could not identify its own likely errors accurately.

3.1. Correction Types

Of the correction types we labeled, the largest number were REPs and OMITs, as shown in Table 1, which shows over-all distribution of correction types, and distributions for each type of system failure corrected. Table 1 shows that 40% of TOOT corrections were

Table 1: Distribution of Correction Types

	ADD	ADD/OMIT	OMIT	PAR	REP
All	8%	2%	32%	19%	39%
Post-Mrec	7%	3%	40%	18%	32%
Post-Rej	6%	0%	7%	28%	59%

simple repetitions of the previously misrecognized turn. While this strategy is often suboptimal in correcting ASR errors [7], REPs (45% error) and OMITs (52% error) were better recognized than ADDs (90% error) and PARs (72% error).

What the user was correcting also influenced the type of correction chosen. Table 1 shows that corrections of misrecognitions (Post-Mrec) were more likely to omit information present in the original turn (OMITs), while corrections of rejections (Post-Rej) were more likely to be simple repetitions. The latter finding is not surprising, since the rejection message for tasks was always a close paraphrase of "Sorry, I can't understand you. Can you please repeat your utterance?" However, it does suggest the surprising power of system directions, and how important it is to craft prompts to favor the type of correction most easily recognized by the system.

3.2. Prosodic Features of Corrections

In part to test the hypothesis that corrections tend to be hyperarticulated (slower and louder speech which contains wider pitch excursions and more internal silence), we examined the following features for each user turn: maximum and mean fundamental frequency values (F0 Max, F0 Mean); maximum and mean energy values (RMS Max, RMS Mean); total duration; length of pause preceding the turn (Prior Pause); speaking rate (Tempo); and amount of silence within the turn (% Silence).[1] F0 and RMS values, representing measures of pitch excursion and loudness, were calculated from the output of Entropic Research Laboratory's pitch tracker, *get_f0*, with no post-correction. Timing variation was represented by four features: Duration of turn and length of pause between turns was hand labeled. Speaking rate was approximated in terms of syllables in the recognized string per second. % Silence was defined as the percentage of zero frames in the turn, i.e., roughly the percentage of time within the turn that the speaker was silent.

[1] While the features were automatically computed, turn beginnings and endings were hand segmented in dialogue-level speech files, as the turn-level files created by TOOT were not available. Because of some system/user overlap in the recordings, we were able to calculate prosodic features for only 1975 user turns.

To ensure that our results were speaker independent, we calculated mean values for each speaker's corrections and non-corrections for every feature. Then, for each feature, we created vectors of speaker means for recognized and misrecognized turns and performed paired t-tests on the paired vectors. For example, for the feature "F0 max", we calculated mean maxima for corrections turns and for non-corrections for each of our thirty-nine speakers. We then performed a paired t-test on these thirty-nine pairs of means to derive speaker-independent results for differences in F0 maxima between corrections and non-corrections.

Our results provide some explanation for why corrections are more poorly recognized than non-corrections, since they indicate that corrections are indeed characterized by prosodic features associated with hyperarticulation. Table 2 shows that corrections differ from other turns in that they are longer, louder, higher in pitch excursion, follow longer pauses, and contain less internal silence than non-corrections. All but the latter difference supports the hypothesis that corrections tend to be hyperarticulated.

Table 2: Corrections vs. Non-Corrections by Prosodic Feature

Feature	T-stat	Mean Corr - !Corr	P
*F0 Max	3.79	17.76 Hz	0
F0 Mean	0.23	-4.12 Hz	0.823
*RMS Max	4.88	347.75	0
*RMS Mean	2.57	63.44	0.014
*Duration	6.68	1.16 sec	0
*Prior Pause	2.17	.186 sec	0.036
Tempo	1.78	-0.15 sps	0.246
*% Silence	4.75	-.05%	0
*significant at a 95% confidence level ($p \leq .05$)			

To confirm this hypothesis, two of the authors labeled each turn in the corpus for evidence of perceptual hyperarticulation, following [13]. 52% of corrections in the corpus has some perceptual hyperarticulation, compared with only 12% of other turns. Too, hyperarticulated corrections are more likely to be misrecognized than other corrections (70% misrecognitions vs. 52%). However, it is important to note that the hyperarticulation explanation accounts for only 59% of misrecognized corrections in the corpus. There are still a large number of misrecognized corrections that show no perceptual evidence of hyperarticulation.

In our earlier analysis of prosodic differences between correct and incorrectly recognized turns [8], we also found that misrecognized turns differed from correctly recognized turns in f0, loudness, duration, and timing — all features associated with hyperarticulation. And more misrecognitions are hyperarticulated than are correctly recognized turns. But when we excluded perceptually hyperarticulated turns from our prosodic analysis, we found that misrecognized turns were still prosodically different from correctly recognized turns, in the same ways. We hypothesized there that misrecognitions might exhibit tendencies toward hyperarticulation that are imperceptible to human listeners, but not to ASR engines. The same may also be true of non-hyperarticulated, but still prosodically distinct corrections. When we exclude hyperarticulated utterances from our corpus and re-analyze prosodic features of corrections vs. non-corrections, we find significant differences in duration, rms maximum, rms mean, tempo, and amount of turn-internal silence as we did with the corpus as a whole. So, again, even when corrections are not perceptibly hyperarticulated, they share some acoustic tendencies with turns that are.

3.3. Correction Chains

As noted above, corrections in the TOOT corpus often take the form of chains of corrections of a single original error. Looking back at Figure 1, for example, we see two chains of corrections: In the first, which begins with the misrecognition of turn 776 ("Um, tomorrow"), the user repeats the original phrase and then provides a paraphrase ("Saturday"), which is correctly recognized. In the second, beginning with turn 780, the time of departure is misrecognized. The user omits some information ("A.M.") in turn 781, but without success; an ADD correction follows, with the previously omitted information restored, in turn 783.

Distance of a correction from the original misrecognized turn — whether calculated as position in chain (e.g. "Saturday" in Figure 1 is the second in the chain correcting turn 776) or further in number of turns from that original error (e.g. "Saturday" here is also 2 turns from the original error), correlates significantly with prosodic variation. An analysis of the relationship between both distance measures and our prosodic features (using Pearson's product moment correlation) shows significant correlations of distance in chain or from original error with f0 maximum ($r=.20, p<.001$; $r=.21, p<.001$) and mean ($r=.27, p<.001$; $r=.29$, $p<.001$), rms maximum ($r=-.09$, $p<.02$; $r=-.12, p<.005$) and mean ($r=-.12, p<.0025$; $r=-.16, p<.001$), absolute duration ($r=.14, p<0$; $r=.16, p<.001$) and duration in number of words ($r=.11, p<.01$; $r=.12, p<.005$), length of preceding pause ($r=.11, p<.005$; $r=.10, p<.01$), and speaking rate ($r=-.05, p<.01$; $r=-.10, p<.02$). The more distant a correction is, in short, the higher it is in pitch, the softer it is, the longer it is, the greater is its preceding pause, and the more slowly it is spoken. In addition, more distant corrections are also more likely to be misrecognized; for distance in turns there is a (negative) significant correlation for concept accuracy ($r=-.13, p<.001$), while both word and concept accuracy decline significantly by position in chain ($r=-.08, p<.05$; $r=-.15, p<.001$). And final corrections in a chain are significantly ($t=4.41, df=38, p<.001$) more likely to be misrecognized than first corrections (47% vs. 41%). So, the more times speakers try to correct an error, the less likely they are to succeed, perhaps because their prosodic behavior changes in ways that do not help the speech recognizer. Curiously, however, our perceptual measure of hyperarticulation is not significantly correlated with either of this distance measures. So, although speakers modify their speech in ways generally consistent with hyperarticulation, their corrections do not necessarily become more hyperarticulated as their attempts to correct continue. Another curious finding is that corrections that are more distant from the turn they immediately correct (e.g. in Figure 1 turn 783 is more distant from the turn it corrects (781) than turn 781 is from the turn *it* corrects, which is 780) tend to be *more* accurately recognized than turns which are closer. Yet prosodically these turns are very like distant turns in a chain or from the original error, being higher in f0 maximum and mean, lower in rms maximum and mean, and longer in seconds and number of words. So in the one case these prosodic changes might be thought to lead to recognition error, where in the other they occur with better recognized corrections.

3.4. Variation by Dialogue Strategy

Dialogue strategy clearly affects the type of correction users make and whether it is successful or not. For example, users more frequently repeat their misrecognized utterance in the SystemExplicit (75% of corrections are repetitions) condition, than in the Mixed-Implicit or UserNoConfirm (both 37% REP); the latter conditions have larger proportions of OMITs and paraphrases. Perhaps this disparity is partly explained by the larger proportion of corrections that follow rejections in the SystemExplicit condition (39% vs. 22% and 19%). Over all, SystemExplicit turns are rejected 6% of the time, while the other conditions have about 10% rejections. Table 3 shows differences in mean length of tasks, number of corrections, number of misrecognitions, and number of misrecognized corrections by dialogue strategy. The fewer misrecognitions,

Table 3: Corrections by System Strategy

Means per Task	System Explicit	Mixed Implicit	User NoConfirm
# Turns	13.4	11.7	16.2
# Corrs	1.3	4.6	7.1
# Misrecs	2.8	6.4	9.4
# Misrec'd Corrs	0.3	3.2	4.8

corrections and misrecognized corrections per task in the SystemExplicit condition may well explain user ratings of the various systems (non-adapt) in the original experiments[9]: When asked to say whether they would be likely to use such a system in future, on a 1-5 scale, subjects scored SE 3.5, MI 2.6, and UNC 1.7. User satisfaction scores were similar: where 40 is the highest scores, users gave SE 31.25, MI 24, and UNC 22.1. So, SystemExplicit is preferred by users, even though MixedImplicit on average takes fewer turns to accomplish a task, suggesting that the large number of misrecognitions and consequent need for correction has a large impact on user preferences. This is consistent with performance functions derived from evaluations of TOOT [9].

Perhaps because correction chains often end unsuccessfully, users frequently "restart" a task within a session. Most restarts occurred in the MixedImplicit and UserNoConfirm conditions and were rarely successful. In non-adaptive tasks, 42% of corrections in the MixedImplicit condition were restarts and 31% in the UserNoConfirm, while none occurred in the SystemExplicit condition. Restarts were misrecognized 77% of the time, compared to 65% of first turns in task. They thus seem to have been a worse strategy than initiating a new task and might prove a useful diagnostic for changing system strategy — or summoning a human operator.

4. DISCUSSION

In this paper we have presented results of an analysis of corrections in the TOOT spoken dialogue corpus. Corrections in our corpus are a serious problem for ASR, being recognized much more poorly than non-corrections but not being rejected any more frequently. Some corrections types are more difficult to handle for systems than others, with repetitions and corrections omitting information from the original turn much better recognized than corrections that add or paraphrase information. Confirming previous studies of repetition corrections, we found that corrections in general differ from non-corrections prosodically: they are higher in f0, softer, longer, follow longer pauses, and contain less internal silence than non-corrections. Also, corrections more distant from the error they are correcting are louder, higher in pitch, longer, slower, and follow longer pauses than closer corrections. Both findings suggest a correlation between corrections and hyperarticulation; however, most prosodic differences persist even when perceptually hyperarticulated turns are removed from the sample, and perceptual hyperarticulation is not significantly correlated with distance from original error. We hypothesize that recognizers may be more sensitive to hyperarticulatory tendencies than humans.

5. REFERENCES

1. L. Bell and J. Gustafson. Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. *Procs. IcPhS-99.*
2. J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. *Procs. ASRU-99.*
3. J. B. Hirschberg, D. J. Litman, and M. Swerts. Generalizing prosodic prediction of speech recognition errors. *Procs. ICSLP-2000.*
4. C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. Evaluating spoken dialog systems for telecommunication services. *Procs. EUROSPEECH-97.*
5. E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Error spotting in human-machine interactions. *Procs. EUROSPEECH-99.*
6. W. J. M. Levelt and A. Cutler. Prosodic marking in speech repair. *Journal of Semantics*, 2, 1983.
7. G.-A. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. *Procs. COLING/ACL-98.*
8. D. J. Litman, J. B. Hirschberg, and M. Swerts. Predicting automatic speech recognition performance using prosodic cues. *Procs. NAACL-00.*
9. D. J. Litman and S. Pan. Empirically evaluating an adaptable spoken dialogue system. *Procs. the 7th International Conference on User Modeling*, 1999.
10. C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *JASA*, 1994.
11. S. L. Oviatt, G. Levow, M. MacEarchern, and K. Kuhn. Modeling hyperarticulate speech during human-computer error resolution. *Procs. ICSLP-96.*
12. Soltau, H. and A. Waibel. Specialized Acoustic Models for Hyperarticulated Speech. *Procs. ICASSP 2000.*
13. E. Wade, E. E. Shriberg, and P. J. Price. User behaviors affecting speech recognition. *Procs. ICSLP-92.*