UNDERSTANDING RECOGNITION FAILURES IN SPOKEN CORRECTIONS IN HUMAN-COMPUTER DIALOG

Gina-Anne Levow

University of Maryland

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. Here we focus on those adaptations - phonological and durational - that are most likely to adversely impact the accuracy of speech recognizers and serve to explain the observed decrease in recognition accuracy on spoken corrections. We identify serveral phonological shifts from conversational to clear speech style. In addition, we compare the observed durations of user utterances from the field trial to those predicted by a speech recognizer's underlying model. We determine that while words in all positions may increase in duration in spoken corrections, those in final position are significantly more strongly affected than those in non-final position. Furthermore, we find that divergence from predicted duration was more marked in corrections of misrecognition errors than for those in corrections of rejection errors. These systematic changes argue for a general hierarchical model of pronunciation and duration, that extends beyond the word or sentence level to incorporate higher-level features from discourse or dialogue.

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, pitch, and pronunciation. We turn to a comparison of observed durations and pronunciations in both original inputs and repeat corrections with those predicted by a speech recognizer's underlying model. The systematic divergences serve both to explain the increase in recognition errors on spoken corrections and argue for a hierarchical model of duration that incorporates information from dialogue structure.

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 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. It also elucidates the divergences between observed user utterance durations in original inputs and repeat corrections relative to those predicted by a typical speech recognizer durational model. 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, & 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 [11]. 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, [1] and [4] 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 [6] and [9], 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, [8] and [10], 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. In previous work [5], we demonstrated that the significant differences between original inputs and spoken corrections in a conversational spoken language system, in duration, pitch, and pause features, could, in turn, be used to train a decision tree classifier to distinguish between original inputs and spoken corrections, of both misrecognition and rejection errors.

In the area of durational modelling, [2] presented a hierarchical durational model that captures differences related to position in stressed or unstressed syllable, word, and phrase. This more fine-grained model improves prediction of phoneme durations.

The current paper extends work in these areas by examining how, given an identification of utterance as a correction, one could adapt the speech recognizer model to compensate for the systematic differences between original inputs and repeat corrections.

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.

Following [8], [9], and [7], we coded a set of acoustic-prosodic features to describe the utterances. These features fall into four main groups: durational, pause, pitch, and amplitude. For this discussion we focus on durational measures.

5.1. Duration

The values for duration are 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 [3]. 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.

6. MOTIVATION

We observed in the initial discussion of the collected data that there was a large disparity between the probability of a recognition error occurring after a correct recognition and the probability of an error following a failed recognition, This contrast leads to the phenomenon of error "spirals", in which multiple successive recognition errors arise. These error spirals are particularly frustrating for users; user evaluations of spoken language systems have been shown to be closely tied to the ease or difficulty of correcting recognition errors. In a study of correction strategies in which extended error spirals were simulated [8], the designers of the study found it necessary to scale back the maximum spiral depth (number of successive failures) to six, from an original depth of ten, when even pilot subjects became so frustrated that they refused to complete the tasks.

USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8,
up 2 3/8 since yesterday.
USER: Give me the price for AT&T.
SYSTEM: Data General is at 63 1/4, up 1/3.
USER: Give me the price for AT&T.
SYSTEM: Hewlett-Packard is at 83 3/8, up 2 3/8.
USER: Give me the price for
American Telephone and Telegraph.
SYSTEM: Excuse me?

In the SpeechActs field trial, error spirals proved to be a common problem for users. One subject encountered a sequence of 15 consecutive recognition failures, to each of which he responded with another attempt at the same utterance, showing remarkable persistence. In fact more errors occurred within the spirals than in first repeat correction position. Clearly, the adaptations that users employ to correct recognition errors in many cases seem to yield the opposite result.

As we observed in [5], these adaptations can be used to identify the corrective force of these utterances, which could not be recognized solely by lexical marking or repetition of lexical content. Clearly these changes provide useful and necessary information to properly interpret the user's intent in uttering the sentence. We argue that it is, in fact undesirable to train users to avoid these adaptations; it is also difficult to do so. Users are often opaque to system directions; a classic example is the oft-reported difficulty of eliciting a simple "yes" or "no" response from a user, even when the user is explicitly prompted to do so. However, just as we note the utility of these cues for interpreting the corrective force of the utterance, we must recognize the severe negative impact that they have on speech recognizer performance. We will demonstrate that these systematic adaptations have specific implications for the design of speech recognizers that will be more robust to the types of changes characteristic of correction utterances.

7. DURATION-RELATED CHANGES

In previous work on spoken corrections ([8], [5]), we noted three classes of systematic changes between original input and repeat correction utterances. There were (1) significant increases in duration, (2) increases in pause measures, and (3) significant decreases in utterance-wide normalized pitch minimum. Most contemporary speech recognizers strip out and normalize for changes in pitch and amplitude; thus pitch and amplitude effects are less likely to have a direct impact on recognizer performance, though pitch features do prove useful in identifying correction utterances. Thus, in this discussion, we will focus on effects of duration and pause changes that can impact recognition accuracy by causing the actual pronunciation of correction utterances to diverge from the speaking models underlying the recognizer.

7.1. Phonetic and Phonological Changes

One of the basic components of a speech recognizer is a lexicon, mapping from an underlying word or letter sequence to one or more possible pronunciations. In conjunction with a grammar, this lexicon constrains possible word sequences to those that the recognizer can identify as legal utterances. There is a constant tension in speech recognizer design between creating the most tightly constrained language model to improve recognition accuracy of those utterances covered by the model and creating a broader-coverage language model to allow a wider range of utterances to be accepted but increasing the perplexity of the model and the possibility of misrecognitions.

In addition to examining the suprasegmental features, we also examined phonological contrasts between original inputs and repeat corrections. We found that more than a third of the original-repeat pairs exhibited some form of phonological contrast, to various extents.

In most of these phonological changes, we found contrasts between the classic dictionary or citation form of pronunciation of the utterance, usually in the repeat correction, and a reduced, casual, or conversational articulation most often in the original input. These changes can be viewed as shifts along a conversational-to-clear-speech continuum [8]. Some examples illustrate these contrasts. Consider, for instance, the utterance "Switch to calendar." The preposition 'to' is a common function word, and this class of words is usually unstressed or destressed and surfaces with a reduced vowel as 'tschwa', even though the citation form is 'too'. Similar reductions are found with a variety of function words, e.g. 'the' which usually appears as 'th schwa' or 'a' as 'schwa'. Throughout the data set of original-repeat pairs we find more than 20 instances of a shift from reduced vowels, surfacing as 'schwa's in the original input utterances, to unreduced and occasionally stressed vowels in the repeat correction utterances. Some instances involve extreme lengthening often accompanied by oscillation in pitch similar to a calling pitch contour [6]. A typical example would be the word 'goodbye' that surfaces as 'goodba-aye'. Approximately 24 instances of this type of lengthening occurred in the data.

These reduced-unreduced contrasts are not limited to vowel instances; a similar phenomenon takes place with released and aspirated consonants. For instance, 't' in the word 'twenty' can fall anywhere along a continuum from essentially elided (unreleased) 'tweny' to flapped 'twendy' to the released and aspirated of citation form 'twenty'. These contrasts are also frequent in SpeechActs data, as in 'nex' in an original input becoming 'next' in a repeat correction, or the frequent elision of the 'd' in goodbye, most often in original inputs.

7.2. Durational Modeling

The conversational-to-clear speech contrasts discussed above are all phonetic and phonological changes which derive from a slower, more deliberate speaking style. In this section we will discuss how increases in duration and pause ([8], [5]) play out in terms of differences between observed utterance durations and speech recognizer model mean durations. We will demonstrate large, systematic differences between observed and predicted durations. This disparity is a cause for concern in speech recognition. In scoring a recognition hypothesis, two measures play significant roles: the score of the frame feature vector as a match to the model feature vector of the speech segment, and a timing score penalty assessed on phonemes that are too long or too short in the Viterbi decoding stage. In other words, recognition hypotheses will be penalized based on the amount the observed duration exceeds the expected duration. We will show that such a mismatch arises for a majority of the words in correction utterances and greater than two-thirds of the words in final position in correction utterances, where correction and phrase-final lengthening effects combine.

We obtained mean durations and standard deviations for a variety of phonemes [2]. For each word in the SpeechActs data set we computed a mean measure of predicted duration by summing the corresponding mean durations for each phoneme in the word. These mean duration measures were then compared to the observed word durations in each of the original input and repeat correction utterances in the data set. [2]

and repeat utterances as shifts from model duration in terms of number of standard deviations from the mean.

7.3. Overall Model Mismatches

The first figure above (Figure 1) presents histograms for all words and for all correction types, with the originals as a thick line and the corrections as a thin line. . Each point on the x-axis is one-half standard deviation. Note, there are very few instances of words less than the mean and also none less than a standard deviation below the mean. There is a large peak for the durations just slightly above the mean, corresponding to values between the mean and one-fourth standard deviation above the mean. The remainder of the words, approximately one-half for all correction types, exceed the mean by at least a standard deviation. The mean value for words in original inputs is 1.0987 standard deviations above the model mean; the median is at 0.8678. In contrast, for correction utterances, the observed mean rises to 1.353 standard

[2] The durations of a small number of words with initial unvoiced stops may have been affected by the conservative approach to marking initial closure, used for pause scoring.

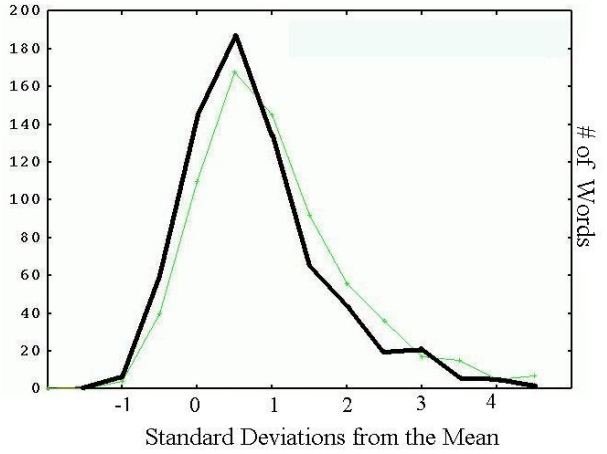

Figure 1: Overlapping Histograms: All Correction Types: Original (thick line) and Correction (thin line): Histogram of Word Duration Shifts from the Mean, in Standard Deviations.

deviations above the mean; with the median value at 1.0750. This shift represents a significant increase in durations. ($t = 3.6$, $df = 1398$, $p < 0.0005$).

The above figures raise the following question: what is the source of this difference from the model durations? It is clearly exacerbated for the repeat corrections, but it is also very much present for words in original inputs as well. Is it simply that the TIMIT durations are a terrible match for conversational, SpeechActs utterances? Or is there a more general explanation for the problem?

7.4. Contrasts by Sentence Position

To answer these questions, we further divide the word duration data into two new groups: words in last position in an utterance and all other words. Phonology argues that phrase- and utterance- final regions undergo a process referred to as phrase-final lengthening, which increases durations in words preceding phrase boundaries. In fact, one of the goals of [2] was to identify meta-features, such as phrase finality, that might change the expected duration of phonemes.

First we look at histograms contrasting shifts from the mean duration for original inputs and repeat corrections for words in non-final position. Graphs for words from all correction types (Figure 2) and corrections of misrecognitions only (Figure 3) are shown below. These figures contrast strongly with the distributions for all words. Instead, the distribution has a single large peak and two fairly narrow tails. In fact, these durations appear to be in closer agreement with the model, aside from having a slightly higher average duration with most durations falling between the mean and one-quarter of a standard deviation above the mean. The observed means for original inputs in non-final position are 0.7894 and 0.5520, and medians at 0.6404 and 0.4348, for all correction types and corrections of misrecognitions only, respectively, closer to the expected duration model. Secondly, we should note the difference between the distribution for words in original inputs and for words in repeat corrections, for non-final positions. The positions of the highest and second highest peaks reverse, placing the largest peak for correction utterances at approximately one-half standard deviation above the mean. Quantitatively the contrast

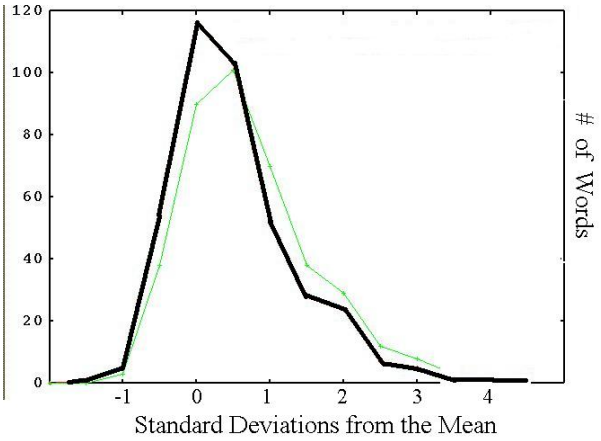

Figure 2: Overlapping Histograms: All Correction Types: Non-Final Words Original (thick line) vs Corrections (thin line) Duration Distribution

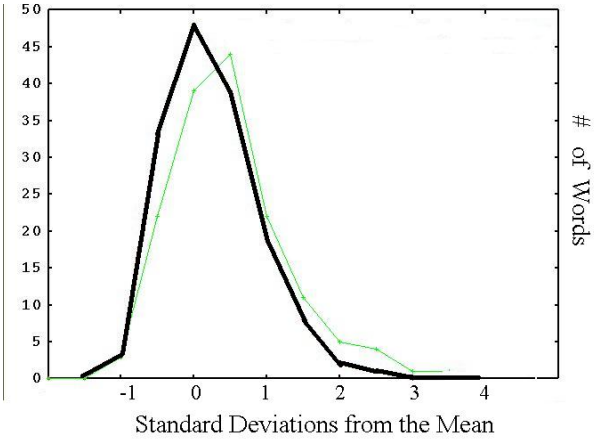

Figure 3: Overlapping Histograms: Corrections of Misrecogntions: Non-Final Words Original (thick line) vs Corrections (thin line) Duration Distribution

between original and repeat inputs is even more apparent. The means rise from 0.7894 to 1.0556 for corrections of all types, and from 0.5520 to 0.7565 for corrections of misrecognition errors. These increases reach significance for corrections of all types (T-test: two-tailed, t= 3.3, df = 792, $p < 0.005$), and approach significance for corrections of misrecognition errors (T-test, two-tailed: t = 1.65, df = 204, p = 0.0518).

Now we examine only those words in utterance-final position, again displaying overlapping histograms for the distribution of durations for original inputs and repeat corrections. Again we observe strong contrasts with the preceding figures. As suggested by phonological theory and [2]'s analysis, there is a significant increase in duration of words in final position relative to a general mean duration. Instead of a large peak less than one-quarter of a standard deviation above the mean, the largest peak for original inputs has shifted to between one-half and three-quarters of a standard deviation above the mean, depending on the error type. Not only is there a shift for the original inputs, but the words drawn from the repeat corrections shift even further.

Shifting to a more quantitative analysis, we find that the mean

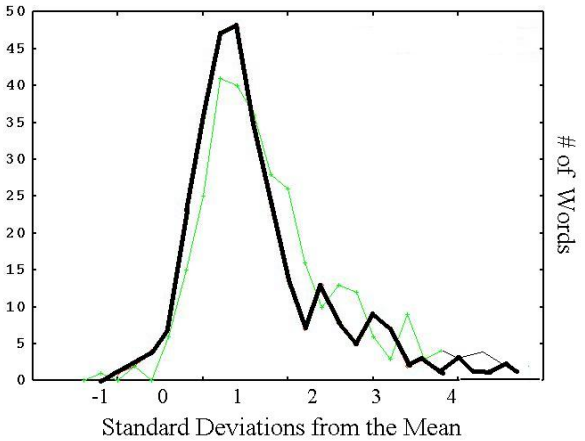

Figure 4: Overlapping Histograms: All Correction Types: Final Words Only Original (thick line) vs. Correction (thin line) Duration Distribution

value for words in final position in original utterances is double the value for words in non-final positions. A similar relationship holds for repeat corrections, with corrections of misrecognition errors experiencing a greater increase.

Correction Type	Repeat?	Non-final	Final
All Types	Original	0.7894	1.5039
All Types	Repeat	1.0556	1.7446
Misrecognitions	Original	0.5520	1.1358
Misrecognitions	Repeat	0.7565	1.514

All of these contrasts between words in final and non-final positions are highly significant. (T-test: two-tailed, $p < 0.001$) These two groups should thus be viewed as coming from different distributions. The largest portion of the durational contrast between original inputs and repeat corrections arises from further increases in duration to the already lengthened words in phrase-final position.

The first graph below (Figure 4) illustrates the distributions for utterance-final word durations for corrections of all error types. The second graph (Figure 5) illustrates the analogous distribution for corrections of misrecognitions errors alone. We observe not only an overall rightward shift in the distributions for all repeat corrections in contrast to original inputs, but also a difference between the two groups of corrections. While the highest peak for corrections of all types decreases in amplitude with more 66% of words exceeding the mean by more than one standard deviation, the change for corrections of misrecognition errors is even more dramatic. The position of the highest peak actually increases by one-quarter of a standard deviation moving the distribution closer to a normal distribution (kurtosis = 3.0883, skewness = 0.4759, the lowest such measures for all distributions), centered now at one standard deviation above the expected mean. Both of these increases from original to repeat correction are shown to be significant. (T-test: two-tailed, t = 2.07, df = 604, $p < 0.02$ for corrections of all types and t = 2.73, df = 174, $p < 0.005$ for corrections of misrecognitions only).

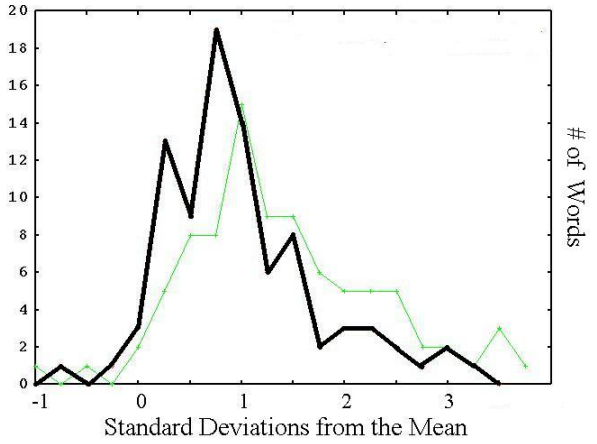

Figure 5: Overlapping Histograms: Corrections of Misrecognitions: Final Words Only Original (thick line) vs. Correction (thin line) Duration Distribution

8. SUMMARY

This more detailed analysis of distribution of word durations in original inputs and repeat corrections allows us to construct a more unified picture of durational change. Basic duration models hold fairly well for pre-final words in original inputs, and show an increase to between one-fourth and one-half standard deviation above the mean in repeat corrections. In contrast, utterance-final words are very poorly described by these models. In all utterances the final words are subject to the effects of phrase-final lengthening, causing them to deviate from the models which suffice for other positions within the utterance. In addition, the effects of corrective adaptations, in turn, interact with and are amplified by the effects of phrase final lengthening. These combined effects cause words in utterance- final position of repeat corrections to deviate most dramatically from models of duration that do not take these effects into account. We see that these changes are most evident in corrections of misrecognition errors where a contrast with basic speaking style is most needed to inform the listener of corrective intent, in the absence of cues available for corrections of rejection errors where the system itself is aware of the recognition failure. Finally, the dramatic changes to utterance-final duration under the dual effects of phrase-final lengthening and corrective adaptation indicate the need for a durational model for speech recognition that can take this meta-information, such as position in utterance and discourse function, into account and further provide a starting point for the implementation of such a model.

9. CONCLUSION

The changes in duration that we observed in this acoustic analysis reflect not only a contrast between original inputs and repeat corrections but a shift away from the models underlying a speech recognizer. Phonological changes from reduced to citation form, following a conversational- to-clear speech continuum, move counter to the painstakingly modeled co-articulation effects of conversational speech. The presence of corrective speech acts signals the need for a different model of phoneme duration to prevent error spirals. Prosodic features can also be used to identify such dialogue states. This analysis of durational and phonological change in spoken corrections demonstrates the importance of understanding and modeling the interaction of dialogue structure and prosody.

10. REFERENCES

1. J. Bear, J. Dowding, and E. Shriberg. 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, 1992.
2. Grace Chung. Hierarchical duration modelling for speech recognition. Master's thesis, Massachusetts Institute of Technology, 1997.
3. D. Colton. Course manual for CSE 553 speech recognition laboratory. Technical Report CSLU-007-95, Center for Spoken Language Understanding, Oregon Graduate Institute, July 1995.
4. P.A. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proceedings of the ACL*, pages 295–302, New Mexico State University, Las Cruces, NM, 1994.
5. Gina-Anne Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of COLING-ACL '98*, 1998.
6. C.H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustic Society of America*, 95(3):1603–1616, 1994.
7. 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. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proceedings of the International Conference on Spoken Language Processing*, 1996. supplementary paper.
8. S.L. Oviatt, G. Levow, M. MacEarchern, and K. Kuhn. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 801–804, University of Delaware and A.I. duPont Instit., 1996.
9. E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Eurospeech '97*, 1997.
10. M. Swerts and M. Ostendorf. Discourse prosody in human-machine interactions. In *Proceedings of the ECSA Tutorial and Research Workshop on Spoken Dialog Systems - Theories and Applications*, June 1995.
11. N. Yankelovich, G. Levow, and M. Marx. Designing SpeechActs: Issues in speech user interfaces. In *CHI '95 Conference on Human Factors in Computing Systems*, Denver, CO, May 1995.