ELSEVIER

Speech Communication 36 (2002) 147–163

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Adaptations in spoken corrections: Implications for models of conversational speech

Gina-Anne Levow

Institute for Advanced Computer Studies, University of Maryland, College park, MD 20742, USA

Received 30 January 2000; received in revised form 6 November 2000; accepted 29 December 2000

Abstract

Miscommunication in spoken human–computer interaction is unavoidable. Ironically, the user's attempts to repair these miscommunications are even more likely to result in recognition failures, leading to frustrating error "spirals". In this paper we investigate users' adaptations to recognition errors made by a spoken language system and the impact of these adaptations on models for speech recognition. In analyzing over 300 pairs of original and repeat correction utterances, matched on speaker and lexical content, we found overall increases in 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. We identify several phonological shifts from conversational to clear speech style. We determine that the observed durations of spoken user corrections from a field trial represent increases over, and divergences from, those derived from a speech recognizer's underlying model. Furthermore, words in final position diverge significantly more than those in non-final position, due to the additional effects of phrase-final lengthening. These systematic changes argue for a general model of pronunciation and duration, extending beyond the sentence level to incorporate higher-level dialog features, and illustrate important features for such a model to capture.

Keywords: Spoken language systems; Dialogue; Prosody

1. Introduction

1.1. Challenges posed by misrecognitions

Misrecognitions of user commands are unavoidable in spoken language systems. These misrecognitions in turn necessitate corrections by the user in order to accomplish their task. In the system analyzed in this study, approximately 25% of user inputs resulted in some type of recognition failure. Ironically, we observe that these corrections are even more likely to be misrecognized than other inputs. Specifically, given a prior correct recognition, the probability of an error on the next input is 18%; however, following a recognition error, the probability of another error rises to 44% – almost 2.5 times the baseline rate. More formally, if $\Pr(E \mid C) = 0.18$, then $\Pr(E \mid E) = 0.44$.

Not only are corrections common and problematic for recognition, they can have a strong adverse effect on user assessments of system quality. Ease of error resolution, rather than absolute recognition accuracy of a system, is closely linked to evaluations of a speech system. Users prefer

E-mail address: gina@umiacs.umd.edu (G.-A. Levow).

0167-6393/02/$ - see front matter
PII: S0167-6393(01)00031-0

systems in which they can easily recover from errors; unfortunately, the decrease in recognition accuracy that accompanies corrections runs directly counter to this need for easy recovery. In addition, it leads to the commonly observed phenomenon of error "spirals", successive errors in which the user repeatedly attempts to enter the same command, but the system continues to blithely misrecognize the input. This experience is particularly frustrating for users, many of whom are remarkably persistent, attempting the same input for as many as 15 times. A brief example of an error "spiral" appears below.

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?

1.2. Difficulties in recognizing corrections

We would like to understand why recognition accuracy decreases for spoken corrections. This task is complicated by several factors. First, it clearly cannot be solely lexical choice that hampers recognition for spoken corrections, since there is actually relatively little variation in utterances. In the corpus studied, 500 lexical strings accounted for 6700 of the observed utterances, almost 80%. Thus we cannot identify corrections based on lexical content or repetition, even if recognized, since repetition is frequent in these interactions. If lexical content does not provide an explanation, we must consider other potential sources of variation. The recognizer itself treats each attempt at recognition independently; it does not change state after a recognition error, which it typically is unable to identify in the first place.

A probable source for these errors is acoustic-prosodic variation, which, we will demonstrate, is a systematic characteristic of spoken corrections. Speech recognition systems often ignore or try to normalize away much of this type of acoustic variation, for instance by normalizing amplitude or pitch across speakers or utterances. However, not all differences can be compensated for in this fashion.

One might suggest trying to change user behavior by training or instructions that help users avoid these types of problematic alterations. However, this solution may not be either practical or desirable. Users are often quite opaque to system direction or correction. An example of this tendency appears in user responses to "yes/no" questions: even when explicitly prompted to "Say yes or no.", users respond in other ways. Furthermore, these acoustic changes, while possibly problematic for automatic speech recognition, are a natural part of human conversation and can actually provide useful cues to the corrective intent of the utterance.

The impact of corrections and correction handling can be understood more readily in a system context as illustrated below. Fig. 1 depicts a basic spoken dialog system pipeline in the upper row, augmented below with components to facilitate

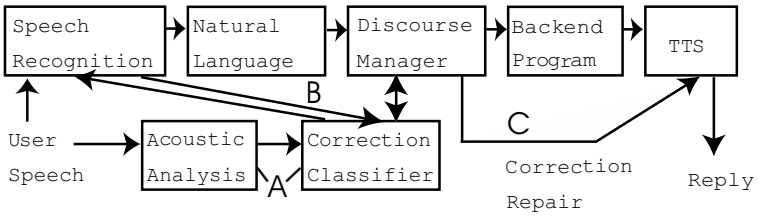

Fig. 1. A simple pipelined dialog system augmented to handle spoken corrections. This figure assumes a black-box recognizer such that additional acoustic analysis and classification must be performed in parallel: (A) identify corrections; (B) context-adaptive recognition; (C) correction repair interaction.

more effective handling of spoken corrections. Adding acoustic analysis and classification can enable identification of spoken corrections (A). This classification could then select a context-adaptive speech recognizer that can compensate for the acoustic adaptations of spoken corrections (B). In addition, identification of a corrective utterance could cause the dialog manager to invoke a repair strategy (C). In the remainder of the paper, we will focus on identifying acoustic adaptations that are likely to impact recognition of spoken corrections and the implications for recognizer design.

1.3. Outline

In this paper, we will introduce some related work on spoken corrections and self-corrections and duration modeling. We will then describe the systematic acoustic variations in spoken corrections by field trial users of a spoken dialog system that we will introduce. We will demonstrate that there are significant differences in duration, pause, and pitch features between original inputs and repeat corrections in lexically matched pairs. We will then examine more closely those variations that are likely to impact recognition accuracy, specifically duration and phonological changes. We will demonstrate that the acoustic correction adaptations we identify in the initial analysis correspond to significant divergences from baseline recognizer models of pronunciation and duration.

2. Related work

Since full voice-in/voice-out spoken language systems are still a relatively recent development, little work has been done on error correction dialogs in this context, though a modest body of work is beginning to emerge. 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, Heeman and Allen (1994) and Bear et al. (1992) 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, this is not the case for most speech recognition systems currently available. Alternative approaches are described in (Nakatani and Hirschberg, 1994; Shriberg et al., 1997) and have emphasized acoustic-prosodic cues, including duration, pitch, and amplitude as discriminating features.

The first studies that focussed on spoken corrections of computer misrecognition (Swerts and Ostendorf, 1995; Oviatt et al., 1998) also found significant effects of duration, and in (Oviatt et al., 1996), pause insertion and lengthening played a role. Subsequent work (Fischer, 1999; Pirker et al., 1999; Bell and Gustafson, 1999) has identified a similar array of acoustic changes for other languages, including German and Swedish. Most of these studies have been Wizard-of-Oz studies rather than trials of implemented speech dialog systems and thus provide no information about the impact of these acoustic characteristics on recognition. In previous work (Levow, 1998), we demonstrated that the significant differences in duration, pause, and pitch features between original inputs and spoken corrections in a conversational spoken language system 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 modeling Chung (1997) 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 and performance in word-spotting experiments.

The current paper extends work in these areas by examining contrasts in acoustic-prosodic features of spoken corrections in a fully operational spoken dialog system, in relatively natural, unconstrained user interactions. Furthermore, we consider the impact of these characteristic adaptations on speech recognition to begin to explain the degradation in recognition accuracy observed. We also identify, where appropriate, contrasts

between durations predicted by phoneme duration models, as they relate to position in utterance and original/repeat status. We illustrate acoustic–phonetic changes associated with corrections whose modeling would form a necessary component of an effective model of human–computer conversational speech, and of its duration in particular.

3. Data collection: system, subjects, and overall analysis

The data used in the experiments and analyses throughout this paper were drawn from recordings of user interactions with the Sun Microsystems SpeechActs system during a limited field trial. This section will provide an introduction to the SpeechActs system itself, a general description of the data obtained in the field trial, a discussion of the identification, coding, and analysis of errors encountered by system users, and a description of the criteria by which the utterance pairs used in the acoustic analysis and recognition experiments were selected.

3.1. SpeechActs description

SpeechActs is a prototype spoken language system designed by Sun Microsystems Laboratories Speech group (1993–1995), where the author spent two summers as a student intern. SpeechActs provides a voice-only interface to a variety of common desktop and information feed services. Specifically, it included e-mail reading and sending, access to one's own and other browsable calendars, weather information, stock quotes, time zone and currency conversions, and a notification system. An example interaction [1] is shown 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 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...

Two significant features distinguish SpeechActs from other spoken language systems. First, since it was designed primarily for use over the telephone to provide ubiquitous access, it is a voice-only system. With the exception of password entry and escape sequences which use touch tone keypad input, all user input is spoken, and all output is through synthesized speech; there are no visual displays for feedback. Speech recognition is performed by BBN's Hark speaker-independent continuous speech recognizer, and synthesis is performed by Centigram's TruVoice text-to-speech system. Locally developed natural language processing and interpretation engines feed the speech recognition results to the appropriate speech application interface for each back end system. The restriction to a single, audio modality requires that the user give all necessary information to the system through speech, and allows our analysis to focus exclusively on those speech cues of lexical, phonetic, and acoustic–prosodic form which the spoken modality provides.

Secondly, SpeechActs was designed to provide a "conversational" interface. A conversational interface can best be understood by what it is not. It is not a fixed command language, it is not a form-based input structure, and it does not have rigid vocabulary or syntax. Instead, a conversational interface hopes to provide both ease of use for novice users and efficiency for more experienced users by allowing them to use language which comes naturally for each individual. In addition, it is easy to combine commands or criteria for requests into a single command for more confident and experienced users (e.g. read the third urgent message) or to simply step through the

[1] Designing SpeechActs: Issues in Speech User Interface Design (Yankelovich et al., 1995, p. 2).

information with a sequence of simple commands for novice users (e.g. "Go to urgent messages", "Next", "Next", "Next"). All new users are provided with a wallet-sized information card with examples of common commands for each application, but users each rapidly develop their own distinct style and vocabulary.

3.2. Data collection and coding

Now that we have provided a general overview of the SpeechActs system, let us turn to a more detailed description of the data collection process. As discussed above, SpeechActs was deployed for a limited field trial over an analog telephone connection, so that it could be accessed from home, office, hotel, or even a busy, noisy airport terminal. All interactions were recorded automatically during the course of the conversation. All speech, both user input and system synthesized responses were digitized and stored at 8 kHz sampling rate in 8-bit mu-law encoding on a single channel, compatible with native system hardware and the limitations of analog telephone lines. In addition to the stored audio, speech recognizer results, natural language analysis results, and the text of all system responses were recorded and time stamped.

Next, all user utterances were textually transcribed by a paid transcriber. Each transcription of user input was paired with the speech recognizer output for that utterance. Each of these pairs was assigned one of four accuracy codes:

- Correct: Recognition and action correct
  User said: Read message one
  System heard: Read message one
- Error minor: Recognition not verbatim; action correct
  User said: Go to the next message
  System heard: Go to UH next message
- Misrecognition: Recognition not verbatim; action incorrect
  User said: Next
  System heard: Fax
- Rejection: No recognition; no action
  User said: Read message one
  System heard: nothing

The use of the "Correct" code should be evident. The "error minor" code assignments generally resulted from a misrecognition of a non-content word (e.g. wrong tense of an auxiliary verb, incorrect article, insertion of "um" or "uh") for which the robust parsing of the natural language component could compensate. The "misrecognition" and "rejection" codes were assigned in those cases where a user could identify a failure in the interaction. Utterances coded either as Misrecognition or Rejection could also receive an additional tag, out-of-vocabulary (OOV). This tag indicates that either words not in the recognizer's vocabulary or constructions not in the system's grammar were used in the utterances. For simplicity, however, we refer to all these cases as OOV. Two examples appear below:

- Unknown word: Rejection
  User said: Abracadabracadabra
  System heard: nothing
- Unknown construction: Misrecognition
  User said: Go to message five eight six
  System heard: Go to message fifty six
  Grammar knows: Go to message five hundred eighty six

In total, there were 7528 recorded user utterances from the field trial. Of these, 4865 were correctly recognized by the speech recognition pass, and 702 contained minor recognition errors, but still resulted in the desired action. There were 1961 complete recognition failures: 1250 of which were rejection errors and 706 of which were substitution misrecognition errors. The remaining five errors were due to system crashes or parsing errors. In other words, almost two-thirds of recognition failures were rejections, about twice the number of misrecognitions. [2] Overall, this results in a 25% error rate.

We also observe, like (Shriberg et al., 1992), that there is a higher probability of a recognition error following an error than following a correct recognition. Specifically, the probability of an error after a correct recognition is approximately 18%

[2] Curiously, this ratio of rejection errors to misrecognition errors is reversed from that most often reported in spoken language systems. The relatively high rate of rejection errors may be attributed to the noisy telephone environments in which this system was most often used.

whereas after a recognition failure it rises to 44%, more than 2.5 times as likely. This contrast is evident in the presence of, often lengthy, error spirals in which multiple errors follow a single initiating error. This contrast in recognition accuracy between original and correction utterances motivates the contrastive analysis which follows and efforts to characterize the changes which mark corrections.

3.3. Original input-repeat correction pairs

For the experiments reported below, 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.
ORIGINAL
SYSTEM HEARD: 'nothing'.
CODE: Rejection.

SYSTEM SAID: Sorry.
USER SAID: Read message four eight nine.
REPEAT
SYSTEM HEARD: "nothing"
CODE: Rejection

In total, there were 303 of these original-repeat pairs: 215 resulting from rejections and 88 from misrecognitions.

4. Acoustic analysis

In the previous section we described in detail the environment in which the human–computer spoken correction data were collected. We explained the selection of 303 original input-repeat correction pairs, of which 88 were corrections of misrecognition errors (hereafter, CMEs) and 215 were corrections of rejection errors (CREs). In this section we will describe a group of acoustic analyses performed on these groups of utterance pairs. Specifically, we analyze these utterances under four broad classes of acoustic–prosodic features: duration, pause, fundamental frequency ($f0$), and amplitude. These measures draw from much of the literature discussed in Section 2, but are based most heavily on those in (Oviatt et al., 1996; Ostendorf et al., 1996). We will demonstrate significant differences between original input and repeat correction utterances in duration, pause, and fundamental frequency.

4.1. Duration

Duration has long been known to play an important role in a wide variety of speech phenomena. Ends of phrases and utterances are characterized by phrase-final lengthening (Allen et al., 1987). [3] Final positions in lists are denoted by increased duration ('t Hart et al., 1990). Stressed and accented syllables are longer than those that are destressed or unstressed (Nooteboom, 1997). [4] Discourse segment-initial utterances also exhibit increases in duration relative to segment-internal utterances (Swerts and Ostendorf, 1995). We will show that duration also plays a significant role in spoken corrections.

For the majority of these analyses, the following technique was used to obtain utterance duration measures. A two-step semi-automatic process was required. First, the waveform and the corresponding utterance that had been segmented from the full conversational log were sent to a forced alignment procedure. The procedure used the Oregon Graduate Institute Center for Spoken Language Understanding (CSLU)

[3] Phrase-final lengthening is a phenomenon in which phoneme durations become elongated at the end of an utterance.

[4] The first syllable in 'teacher' is stressed; the second syllable is unstressed.

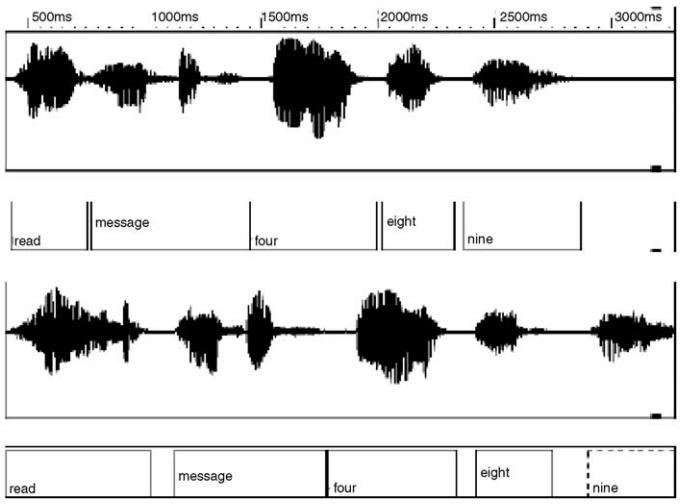

Fig. 2. Original (top) – repeat (bottom) pair with increase in total duration, pause duration, and speech duration.

CSLUsh tools (Colton, 1995) to produce a word-level forced alignment at a 10 ms scale. A second pass over the automatic alignment was performed by a trained analyst. This pass was required to correct for any errors in the original alignment procedure; these errors arose from a variety of factors: background or non-speech noise in the recording, pronunciation mismatch between the aligner dictionary and the spoken utterance, etc. The corrections focussed on three classes of position within the utterance: initial onset of speech, final speech position, and the boundaries of sentence-internal pauses. The goal was to delimit the total duration of the speech in a user turn, rather than to adjust all alignments. We took a conservative approach, only changing an alignment position if there was a better destination position available. From the alignments it was possible to automatically compute the following measures: total utterance duration, total speech duration, total pause duration, total number of pauses, and average length of pause.

4.1.1. Total utterance duration

The first measure we will consider is total utterance duration. Simply put, the total utterance duration is the length in milliseconds from the onset of the user speech in the utterance to the final speech position. Overall, utterances ranged in duration from 210 ms to 5180 ms. An example of an original-repeat pair with increase in total utterance duration appears in Fig. 2.

T-test two-tailed ($t = 1.97,\ df = 604,\ p < 0.05$) indicates a significant increase in total utterance duration from original to correction utterances. Specifically, the mean length of an utterance is 864 ms for original input utterances and 969 ms for repeat correction utterances. This increase corresponds to a 12.15% increase in total utterance duration.

4.1.2. Total speech duration

Total speech duration calculates the difference between total utterance duration and total pause duration. This measure tries to capture the contribution of the speech segment, rather than increases in number or length of pause, to the increase in total utterance duration. In other words, are users simply pausing more, lengthening phonemes, or increasing both pause and phoneme length? An example of an original-repeat correction pair in which speech duration increases with no corresponding increase in pause number or duration appears in Fig. 3.

T-test two-tailed ($t = 2.17,\ df = 604,\ p < 0.05$) indicates an increase in speech duration from

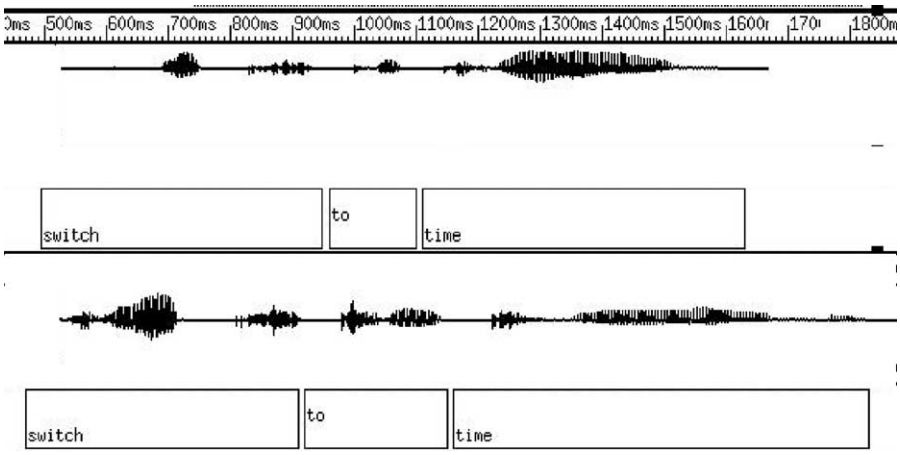

Fig. 3. Original (top) versus repeat (bottom) pair with increase in speech duration only.

original to repeat inputs. This value corresponds to an average increase of 9.5%.

4.2. Pause

Pauses, the presence of unfilled silence regions within utterances, can play a significant role in discourse and utterance-level prosody. In discourse-neutral speech, pauses generally appear at intermediate and intonational phrase boundaries, which often coincide with syntactic phrase or sentence boundaries (Pierrehumbert and Hirschberg, 1990; Bachenko and Fitzpatrick, 1991). Speech systems commonly rely on extended periods of silence, one second or more in length, to identify the end of the user's turn (Yankelovich et al., 1995). While this method is arguably not a good way to detect turn transitions, it is, however, quite effective. The presence of lengthy pauses was found to be a strong cue to the start of a self-repair or other disfluency. [5] (Nakatani and Hirschberg, 1994; Heeman and Allen, 1994; Shriberg et al., 1997) Pauses exceeding 50 ms in length also proved useful in discriminating among speaking styles (Ostendorf et al., 1996).

Here, as noted in the discussion of duration measures, we coded the beginning and ending positions of all pauses in the original-repeat pair data. Silences were coded as pauses only if they exceeded 20 ms in duration. In addition, we excluded all pauses prior to unvoiced plosives (k, t, p) and affricates (e.g. *ch*). [6] This choice was made due to the need to arbitrarily place the starting position of the unvoiced closure for phonemes of these classes, making it impossible to accurately determine the length or even existence of a preceding pause. For each utterance, we then computed the length of each pause, the total number of pauses, and total pause duration. Fig. 4 below illustrates an increase in pause number and duration with little increase in speech duration.

For all pause duration comparisons we considered only those utterances with at least one pause. *T*-test, two-tailed, also yields significant results ($t = 2.2,\ df = 132,\ p < 0.05$) indicating a strong increase in pause duration. Specifically, within utterance silence regions increase from an average of 104 ms for original input utterances to an average of 165 ms, corresponding to an average increase of 59% in total pause duration.

[5] A disfluency is a disruption in normal speech. There are many types: pauses, 'filled pauses', where the speaker inserts 'um' or 'uh', or repetition, as in 'read the the message'.

[6] These phonemes are just a subset of the consonants where the vocal folds do not vibrate at the beginning of the sound. Since speech analysis tools depend heavily on this information, it is hard to identify the start of these sounds precisely.

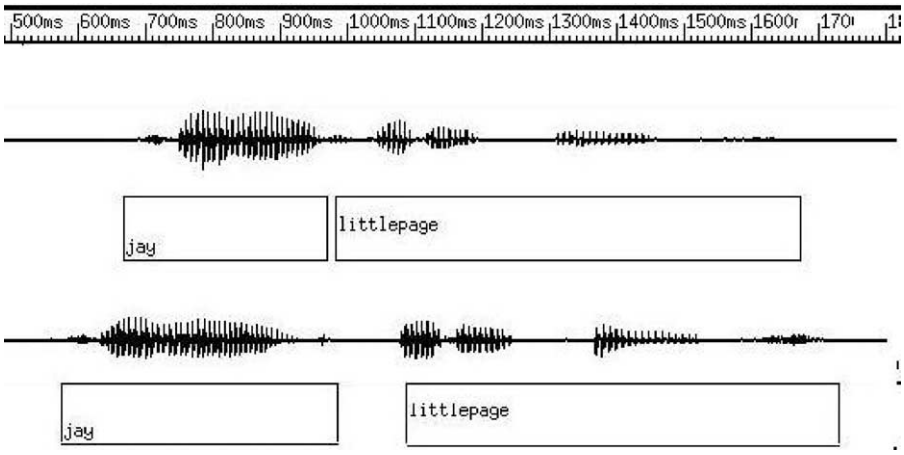

Fig. 4. Original (top) – repeat (bottom) pair with increase in pause duration. Note the insertion of silence between "Jay" and "Littlepage" with no additional increase in word durations.

Total utterance duration was tied to increases in pause duration. To measure these changes we computed the ratio of pause duration to total utterance duration for both original and repeat utterances where pauses occurred. *T*-test two-tailed yielded significant results ($t = 2.28,\ df = 132,\ p < 0.025$) showing an increase in the proportion of silence to total utterance duration. From an average of 7.28% in original utterances, the proportion of silence increases to 10.56%, corresponding to an increase of 46% in the proportion of silence in an utterance.

We computed a final composite measure of speaking rate in the number of syllables per second and normalized by utterance duration. [7] *T*-test two-tailed ($t = 3.6,\ df = 604,\ p < 0.001$) demonstrates a significant *decrease* in speaking rate from original to repeat. The average speaking rate for original utterances was 0.08 dropping to 0.06 for repeat utterances, a decrease of 19.3%.

4.3. Pitch

We also identified significant pitch-related contrasts between original inputs and repeat corrections. These changes are characterized by overall decreases in pitch minimum, specifically decreases in the lowest normalized $f0$ value in the utterance and final word in the utterance, and shifts from question-rise intonation to declarative-fall intonation. Levow (1998) describes these analyses in greater detail.

4.4. Amplitude

Although amplitude is, anecdotally, one of the features commonly associated with corrections, we found that none of the amplitude measures, normalized or not, reached significance.

4.5. Discussion: duration and pause: conversational-to-(hyper) clear speech

We found significant increases in total utterance duration, total speech duration, total pause duration, speaking rate in syllables per second, length per pause, and proportion of silence in utterances between original inputs and repeat corrections. In correction utterances, users speak more slowly both by increasing the duration of phonemes within the utterance and by inserting or lengthening silence regions within the utterance. These changes fit smoothly into an analysis of corrections as shifting from more conversational, casual speech to more clear or careful speech along the

[7] Measuring speaking rate in phonemes or syllables per second is a standard approach in speech systems, compensating for differences in content of the utterances that make measures like number of sentences per second unreliable.

continuum. These changes are consistent with those reported for corrections in Wizard-of-Oz studies by Oviatt et al. (1996) and for conversational versus read speech in (Ostendorf et al., 1996).

This increase in duration seems to be the most robust clear speech attribute. Other types of speech associated with clear style, such as speech to the hearing-impaired or speech to children (motherese), [8] exhibit increases in duration. On one hand, speech to children is often associated with higher pitch and expanded pitch range, while speech to the hearing-impaired lacks these pitch features but is associated with significant increases in loudness (Fernald et al., 1989; Picheny et al., 1986). We have also noted (Levow, 1998) distinctive pitch phenomena associated with corrections which are not shared by other clear speech styles. Speaking rate thus stands out as the most consistent clear speech feature.

While there is a significant increase in pause duration, the increase in pausing is less pronounced for SpeechActs data than in other work. Presence or absence of a 50 ms pause is not a deciding contrast between original and repeat correction as it is for the classes studied by Ostendorf et al. (1996). [9] A likely reason for this observed contrast is the length of the utterances in the current study. In the SpeechActs data overall, the average length of an utterance is between two and three words, and the average analyzed utterance duration is under 2 s. For the SpeechActs data, none of the analyzed utterances exceeded 10 words, while the data in (Oviatt et al., 1996) include 16-digit strings representing credit card numbers. Systems which predict pause location and prosodic phrasing typically use a combination of syntactic phrase structure and number of words or syllables in determining pause placement. Thus pauses are not distributed uniformly over utterances, but are unlikely to appear at all in very brief utterances. The sentences in the SpeechActs data are short enough to discourage pausing, creating this contrast in pause lengths.

[8] Motherese refers to characteristic speech of caretakers to children. It is found in many languages, though more for females than males, and involves expanded pitch range, higher pitch, and longer duration.

[9] Presence of a larger pause duration (70 ms) or larger proportion of silence does play a secondary role in classifying rejection errors with only acoustic information.

5. Implications for speech recognizer design

5.1. 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 (Oviatt et al., 1996), 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.

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 demonstrated in previous work (Levow, 1998), 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 the systematic adaptations of users in the face of recognition errors that have been detailed in the preceding sections have specific implications for the design of speech recognizers that will be more robust to the types of changes characteristic of correction utterances.

5.2. Duration-related changes

In the analysis section, we noted two classes of systematic changes between original input and repeat correction utterances. There were (1) significant increases in duration and (2) increases in pause measures. 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.

5.2.1. 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 of duration, pause, and amplitude discussed in preceding sections, we also examined segmental contrasts between original inputs and repeat corrections. We found that more than a fifth of the original-repeat pairs exhibited some form of segmental contrast, to various extents. Many of these changes occur along what may be called a conversational-to-clear speech continuum, as discussed in (Oviatt et al., 1996).

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. 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 'tə', even though the citation form is 'too'. 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 'tweny' to flapped 'twendy' to the released and aspirated of citation form 'twenty'. These contrasts are frequent in SpeechActs data.

In the contrasts discussed above we observed a shift from a reduced, conversational form in the original input to an unreduced, clear speech form in the repeat correction utterance. We also observed instances of extreme lengthening often accompanied by oscillation in pitch, similar to a calling pitch contour (Nakatani and Hirschberg, 1994). A typical example would be the word 'goodbye' that surfaces as 'goodba-aye'. Approximately 24 instances of this type of insertion occurred in the data between original inputs and repeat corrections.

5.2.2. Durational modeling

The conversational-to-clear speech contrasts and lengthening processes discussed above are all segmental changes which derive from a slower, more deliberate speaking style. In this section we will discuss how the increases in duration and

pause described in the acoustic analysis section 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. These penalties are applied in different ways in different speech recognizers. The CSLU CSLUrp speech recognizer builder directly applies a penalty score for phonemes with duration outside an expected range. Other systems use word duration misalignment in determining the confidence score for a transcription (Ljolje et al., 2000). 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. Furthermore, in a post hoc analysis of errors in the switchboard transcription task, word duration is associated with different types of transcription errors (Greenberg et al., 2000).

We obtained mean durations and standard deviations for a variety of phonemes (Chung, 1997). These durations are normalized durations based on utterances in the ATIS corpus. [10] The Air Travel Information System (ATIS) sentences were collected automatically or semi-automatically through a human–computer spoken language interface to air travel reservation information. As such, the data are a fairly good match with the SpeechActs interaction. However, it is not a perfect match, due to possible differences with controlled, task-based data collection in contrast to the field trial, with microphone versus telephone interactions, and with the types and lengths of utterances elicited in the two contexts – ATIS utterances being longer than typical SpeechActs utterances. [11]

For each word in the SpeechActs data set we computed mean and standard deviation measures of predicted duration by summing the corresponding means or standard deviation of durations for each phoneme in the word. We based word pronunciations on the CMU pronouncing lexicon, applying stressed phoneme duration models to those labeled with primary stress. [12] 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. [13] In addition, we computed the measures separately for words in utterance-final position, where, due to phrase final lengthening and the predominance of content words, we expected durational changes to be at their clearest. We present the durational shifts in original and repeat utterance as shifts from predicted duration in terms of number of standard deviations from the mean. For clarity of display in the figures below, we have binned the number of words according to the number of standard deviations from

[10] We implemented the basic phonemic layer of the model, with some additional phonological form information; however, variability from syllable, word and phonological layers may not be captured here.

[11] The original SpeechActs system was implemented with a binary distribution of the proprietary Hark speech recognizer, tuned for telephone speech. As such, we do not have access to the Hark models themselves and thus select an available model based on the most similar type of interactions.

[12] Function words, e.g. 'the', 'of', generally are assumed to be unstressed, and to take their reduced conversational forms, with reduced vowel forms, chosen based on the dictionary, and reduced vowel and onset lengths. These predictions are still conservative, yielding word durations typically 10–80% longer than average predicted function word durations in highly predictable contexts, as described in (Jurafsky et al., 2001), and 10–50% longer than predictions for words in low predictability contexts, though two or three are shorter. Conversational forms are also predicted for some content words as discussed in the section on phonetic and phonological change above. Some of these factors would be handled as part of the full hierarchical model, but that was not available to us.

[13] 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.

the model mean. Bins are 0.5 standard deviations wide, except where noted, and the graph line is drawn through the center of the bin. So, a point at 0.5 indicates the number of words with a duration between 0.25 and 0.75 standard deviations from the predicted duration.

The first figure below presents distributions for all words and all correction types with the originals in thin lines and the corrections in thick lines. There is a large peak for the durations of corrections at about the mean, increasing over that for original inputs. The remainder of the words, more than one-quarter for all correction types, exceed the mean by at least a standard deviation. The mean value for words in original inputs is 0.15 standard deviations above the predicted mean; the median is slightly below 0. In contrast, for correction utterances, the observed mean rises to 0.47 standard deviations above the mean; with the median value at 0.23. This shift represents a significant increase in durations ($t = 4.59$, $df = 1396$, $p < 0.0001$) (see Fig. 5).

The above figures raise the following question: what is the source of this difference from the predicted 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 there is some mismatch between the ATIS-based speaking-rate neutral durations and SpeechActs utterances? Or is there a more general explanation for the problem?

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. Phonological theory 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 Chung (1997) was to identify meta-features, such as phrase finality, that might change the expected duration of phonemes; that work proposed a technique for handling very long words in pre-pausal position.

First we look at distributions contrasting shifts from the mean duration for original inputs and repeat corrections for words in non-final position. The plot for words from all correction types (Fig. 6) is shown below. This figure contrasts strongly with the distributions for all words.

The observed mean for original inputs in non-final position is −0.14 for all correction types. 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 position of the highest peak shifts one-third of a standard deviation higher. Quantitatively the contrast between original and repeat inputs is even more apparent. The means rise from −0.14 to 0.19 for corrections of all types. These

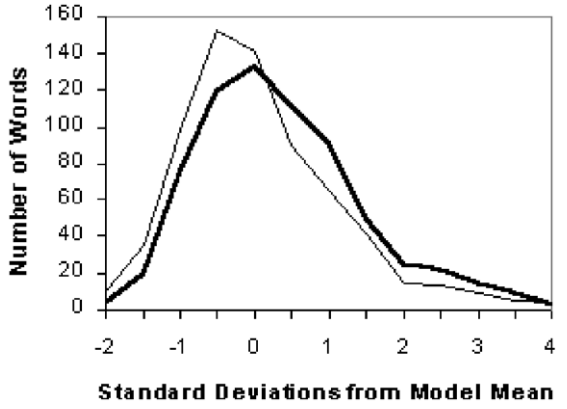

Fig. 5. Overlapping distributions: all correction types: original (thin line) and correction (thick line): word duration shifts from the mean, in standard deviations.

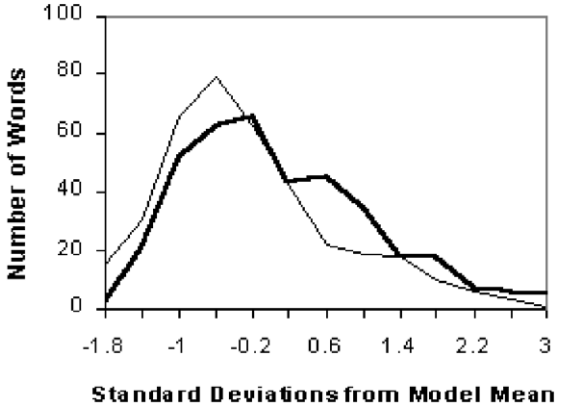

Fig. 6. Overlapping distributions: all correction types: non-final words original (thin line) versus corrections (thick line) durations distribution, bins 0.4 standard deviations.

increases reach significance for corrections of all types (*T*-test: two-tailed, $t = 3.66$, $df = 778$, $p < 0.0005$).

Now we examine only those words in utterance-final position, again displaying overlapping distributions of durations for original inputs and repeat corrections. Fig. 7 illustrates the distributions for utterance-final word durations for corrections of all error types. Fig. 8 illustrates the analogous distribution for corrections of misrecognition 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 than 30% of words exceeding the mean by more than one standard deviation, the change for corrections of misrecognition errors is even more dramatic. The distribution has shifted between one-quarter and one-half of a standard deviation, moving the distribution closer to a normal distribution (kurtosis = 0.75, skewness = 0.86, the lowest such measures for all distributions), centered now at least 0.25 standard deviations above the expected mean. Both of these increases from original to repeat correction are shown to be significant (*T*-test: two-tailed, $t = 3.02$, $df = 604$, $p < 0.003$ for corrections of all types and $t = 2.78$, $df = 174$, $p < 0.0075$ for corrections of misrecognitions only).

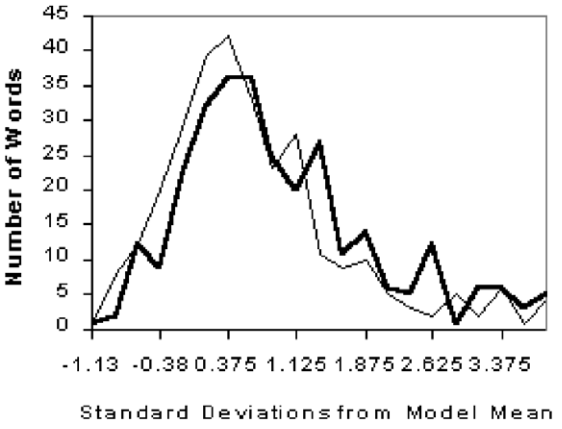

Fig. 7. Overlapping distributions: all correction types: final words only; original (thin line) versus correction (thick line) duration distribution, bins 0.25 standard deviations.

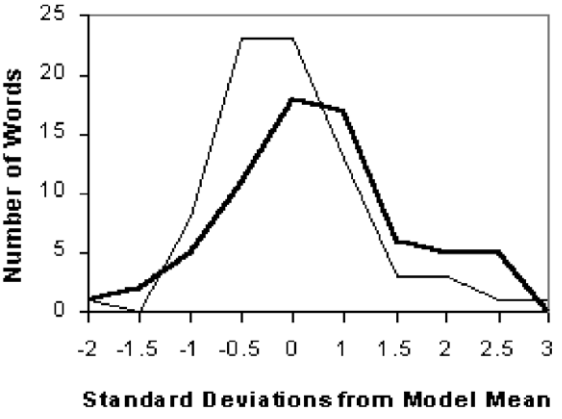

Fig. 8. Overlapping distributions: corrections of misrecognitions: final words only; original (thin line) versus correction (thick line) duration distribution.

Again we observe strong contrasts with distributions of non-final words. As suggested by phonological theory and (Chung, 1997)'s analysis, there is a significant increase in duration of words in final position relative to a predicted mean duration. Instead of a large peak about one-half of a standard deviation below the mean, depending on the error type, the largest peak for original inputs has shifted to at least the mean. 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 value for words in final position in original utterances is 0.65 standard deviations longer than for words in non-final positions. A similar relationship holds for repeat corrections, with corrections of misrecognition errors experiencing a greater increase of 0.87 standard deviations.

Correction type	Repeat?	Non-final	Final
All types	Original	−0.14	0.52
All types	Repeat	0.19	0.83
Misrecognitions	Original	−0.50	0.19
Misrecognitions	Repeat	−0.24	0.68
Rejections	Original	0.09	0.66
Rejections	Repeat	0.47	0.90

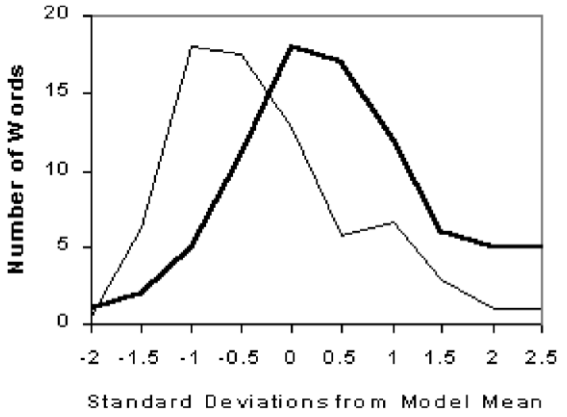

Fig. 9. Overlapping distributions: misrecognition errors: repeat corrections: non-final words (thin lines); final words (thick lines).

All of these contrasts between words in final and non-final positions are highly significant (T-test: two-tailed, $p < 0.0001$). These two groups should thus be viewed as coming from different distributions. [14]

These contrasts between distributions of durations for final and non-final words are clearly illustrated in Fig. 9. Here the plot contrasts final and non-final words for repeat corrections in misrecognition errors. Clearly, words in final position diverge further from predicted durations than those in non-final position, as the table above indicates. Repeat correction status further increases these divergences significantly.

This more detailed analysis 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, though original words in misrecognition errors are fast relative to the model, and show an increase of one-third standard deviation in repeat corrections. In contrast, utterance-final words are relatively 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, add to 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 final versus non-final contrasts are most evident in corrections of misrecognition errors where a contrast with basic speaking style is most needed to inform the system of corrective intent. Finally, the dramatic contrasts for 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.

6. Conclusion

In this paper, we have aimed to better understand communication failure and recovery in spoken human–computer interaction, by examining the acoustic and phonological properties of user utterances in error resolution dialogs. In particular, we have contrasted the characteristics of original inputs and repeat corrections. In field trial interactions with a prototype spoken dialog system, we find an array of systematic contrasts between originals and corrections. There are significant increases in total duration, speech duration, and pause duration. These increases correspond to decreases in measures of speaking rate. Most of these changes can be viewed as shifts toward more careful, "clear" speech.

We observe not only a high rate of recognition failures, approximately 25%, in these spoken interactions, but also an increased rate of errors for spoken corrections. Recognition failures after an

[14] Furthermore, we compute the same statistics excluding the fastest speaker in the cohort by more than 1.75 standard deviations, using mrate (Mirgafiori et al., 1995) to compute speaking rate based on acoustic measures alone. All contrasts remain significant. However, excluding the influence of this subject brings the original inputs in non-final position into very close agreement with the predicted mean, while preserving significant increases from original to repeat and from non-final to final words.

incorrect recognition result occur at more than 2.5 times the rate for utterances following correct recognitions. We assess the relationship between this increase in error rate and the spoken adaptations in user corrections observed in acoustic analysis, by comparing observed durations to those predicted by speech recognizer duration models. We find that the acoustic–prosodic changes reflect not only a contrast between original inputs and repeat corrections but also 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. In addition we observe a very skewed distribution of word durations, that in the change from original to repeat correction increases over the predicted durations, derived from a speech recognizer model.

The analysis of durations and phonological change in "conversational" human–computer dialogs suggests a need for a new understanding of the notion of conversational speech in this context. The model of spoken durations (Chung, 1997) on which we based our predicted durations in these experiments is derived from ATIS, a corpus of human–computer interaction speech. We found that durations of non-final words in original inputs are a bit fast relative to those derived from that speaking-rate normalized model, as some classes of errors have faster than predicted speaking rates. This effect, however, is largely diminished by excluding the fastest speaker. Words in final position and in repeat corrections have significantly greater durations. Divergence from predicted lengths is greatest in final words in repeat corrections, where the effects of corrective adaptations and phrase-final lengthening combine.

Clearly, speaking rate is not uniform within utterances or across utterances by the same speaker. Straightforward adaptation techniques like pitch or vocal tract normalization that apply uniformly to all utterances by the same speaker are likely to be inadequate for the greater variability that must be accounted for in a model of conversational duration. Words take on greater duration in phrase-final position. There is an overall increase in duration in repeat corrections. [15] However, the magnitude of this increase may not be uniform, since there are some differences based on correction type and position in the utterance that have not yet been fully evaluated. Furthermore, the model must accommodate the discrete segmental changes associated with clear speech that can lead to differences in lengthening effects, such as stress on usually unstressed function words as in the segmental changes described above.

In future work, we plan to explore the development of an extended durational model of human–computer speech, that can incorporate the information not only from word position at the sentence level, but also from dialog position and role. This model should capture the systematic durational and phonological clear speech adaptations observed in spoken corrections. This context-adaptive model should be applied when such corrections are detected, either by cue words such as "No, I meant" or by the presence of these same acoustic features. In this manner, we can hope to improve error recovery by deriailing the frustrating cycle of error spirals.

References

Allen, J., Hunicutt, M.S., Klatt, D., 1987. From Text to Speech: The MITalk System. Cambridge University Press, Cambridge.

Bachenko, J., Fitzpatrick, E., 1991. A computational grammar of discourse-neutral prosodic phrasing in English. Computational Linguistics 16, 155–170.

Bear, J., Dowding, J., Shriberg, E., 1992. Integrating multiple knowledge sources for detection and correction of repairs in human–computer dialog. In: Proceedings of the ACL, pp. 56–63.

Bell, L., Gustafson, J., 1999. Repetition and its phonetic realizations: investigating a Swedish database of spontaneous computer directed speech. In: Proceedings of the ICPhS'99.

[15] Training of a separate model for particularly long words in pre-pausal position was raised in (Chung, 1997). That work also explored speaking rate based normalization that might be applicable to some of the correction-related increases in duration described here.

Chung, G., 1997. Hierarchical duration modelling for speech recognition. Master's Thesis, Massachusetts Institute of Technology.

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

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B.D., Fukui, I., 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. Journal of Child Language 16, 477–501.

Fischer, K., 1999. Repeats, reformulations, and emotional speech: evidence for the design of human–computer speech interfaces. In: HCI International '99.

Greenberg, S., Chang, S., Hollenback, J., 2000. Phonetic and lexical dissection of the hub-5 speech recognition evaluation. In: Proceedings of the Hub-5 Speech Recognition Workshop.

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

Jurafsky, D., Bell, A., Gregory, M., Raymond, W., 2001. Effect of language model probability on pronunciation reduction. In: Proceedings of ICASSP-01, Salt Lake City, Utah.

Levow, G.-A., 1998. Characterizing and recognizing spoken corrections in human–computer dialogue. In: Proceedings of the COLING-ACL '98.

Ljolje, A., Hindle, D., Riley, M., Sproat, R., 2000. The AT&T LVCSR-2000 system. In: Proceedings of the Hub-5 Speech Recognition Workshop.

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

Nooteboom, S., 1997. The prosody of speech: melody and rhythm. In: Hardcastle, W.J., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford.

Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Talkin, E.S.D., Waibel, A., Wheatley, B., Zeppenfeld, T., 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.

Oviatt, S., Levow, G., MacEachern, M., Kuhn, K., 1996. Modeling hyperarticulate speech during human–computer error resolution. In: Proceedings of the International Conference on Spoken Language Processing, University of Delaware and A.I. duPont Institute, Vol. 2, pp. 801–804.

Oviatt, S., MacEachern, M., Levow, G., 1998. Predicting hyperarticulate speech during human–computer error resolution. Speech Communication 24 (2), 87–110.

Picheny, M., Durlach, N., Braida, L., 1986. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. Journal of Speech and Hearing Research 29, 434–446.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, P., Morgan, J., Pollack, M. (Eds.), Intentions in Communication. MIT Press, Cambridge, MA, pp. 271–312.

Pirker, H., Loderer, G., Trost, H., 1999. Thus spoke the user to the wizard. In: Eurospeech '99.

Shriberg, E., Wade, E., Price, P., 1992. Human–machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfcation. In: Proceedings of the DARPA Speech and Language Technology Workshop, pp. 49–54.

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

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

't Hart, J., Collier, R., Cohen, A., 1990. A perceptual study of intonation: an experimental phonetic approach to speech theory. Cambridge University Press, Cambridge.

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