



# Prosodic and other cues to speech recognition failures

Julia Hirschberg<sup>a,\*</sup>, Diane Litman<sup>b</sup>, Marc Swerts<sup>c</sup>

<sup>a</sup> Department of Computer Science, Columbia University, 1241 Amsterdam Avenue, M/C 0401, New York, NY 10027, USA

<sup>b</sup> Department of Computer Science, University of Pittsburgh, 210 South Bouquet Street, Pittsburgh, PA 15260, USA, and LRDC, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, PA 15260, USA

<sup>c</sup> Faculty of Arts, Communication & Cognition, University of Tilburg, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands, and CNTS, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Received 5 June 2002; received in revised form 14 March 2003; accepted 8 January 2004

## Abstract

In spoken dialogue systems, it is important for the system to know how likely a speech recognition hypothesis is to be correct, so it can **reject** misrecognized user turns, or, in cases where many errors have occurred, change its interaction strategy or switch the caller to a human attendant. We have identified prosodic features which predict more accurately when a recognition hypothesis contains errors than the acoustic confidence scores traditionally used in automatic speech recognition in spoken dialogue systems. We describe statistical comparisons of features of correctly and incorrectly recognized turns in the TOOT train information corpus and the W99 conference registration corpus, which reveal significant prosodic differences between the two sets of turns. We then present machine learning results showing that the use of prosodic features, alone and in combination with other automatically available features, can predict more accurately whether or not a user turn was correctly recognized, when compared to the use of acoustic confidence scores alone.

© 2004 Published by Elsevier B.V.

*Keywords:* Prosody; Confidence scores; Recognition error

## 1. Introduction

One of the central problems involved in managing the dialogue in most current spoken dialogue systems (SDSs) is how to recover from system error. The automatic speech recognition

(ASR) component of such systems is prone to make mistakes, especially under noisy conditions, or when there is a mismatch between the speech recognizer's training data and the speakers it is called upon to recognize or if the domain vocabulary is large. Users' evaluations of spoken dialogue systems are highly dependent on the number of errors the system makes (Walker et al., 2000a,b; Swerts et al., 2000) and how easy it is for the user and system to correct them. A further complicating factor is how users behave when confronted with system error. After such errors, they often switch to a prosodically 'marked' speaking

\* Corresponding author. Tel.: +1-212-939-7114; fax: +1-212-666-0140.

E-mail addresses: [julia@cs.columbia.edu](mailto:julia@cs.columbia.edu) (J. Hirschberg), [litman@cs.pitt.edu](mailto:litman@cs.pitt.edu) (D. Litman), [m.g.j.swerts@uvt.nl](mailto:m.g.j.swerts@uvt.nl) (M. Swerts).

style—**hyperarticulating** their speech, in an attempt to help the systems recognize them more accurately, e.g., *I said BAL-TI-MORE, not Boston*. While such behavior may be effective in human–human communicative settings, it often leads to still further errors in human–machine interactions, perhaps because such speech differs considerably from the speech most recognizers are trained on. In attempting to improve system recognition, users may thus in fact make it even worse. Another complication is that when system responses reveal false beliefs in implicit verification questions—as when a system’s attempt to verify new information reveals that it has mistaken a previous user input (e.g., *Where do you want to go from Boston* when the user has said she wants to depart from Baltimore). In such cases, users may become quite confused: they are faced with the choice of correcting the misconception or answering the underlying question asked—or doing both at once (Krahmer et al., 2001).

Given that it is impossible to fully prevent ASR errors, and that error levels are likely to remain high as applications become ever more ambitious, it is important for a system to know how likely a speech recognition hypothesis is to be correct. With such information, systems can **reject** (decide that the best ASR hypothesis should be ignored and, usually, prompt for fresh input) speaker turns that are misrecognized—but not prolong the dialogue by rejecting correctly recognized turns, or they can try to recognize the input again in a following loop using a differently trained ASR system. Alternatively, in cases where many errors have occurred, systems might use a correct knowledge of past misrecognitions in deciding to change their interaction strategy or to switch the caller to a human attendant (Litman et al., 1999; Litman and Pan, 1999; Walker et al., 2000a,b).

Traditionally, the decision to reject a recognition hypothesis is based on acoustic confidence score thresholds (based only on acoustic likelihood), which provide some measure of the reliability of the hypothesis; these thresholds are application dependent (Zeljko, 1996). This process often fails, as there is no simple one-to-one mapping between low confidence scores and incorrect recognitions, and the setting of a rejection

threshold is generally a matter of trial and error (Bouwman et al., 1999). This process is also not necessarily appropriate for dialogue systems, where some incorrect recognitions do not necessarily lead to misunderstandings at a conceptual level (e.g., *Show me the trains* recognized as *Show me trains*). SDSs often need to recognize conceptual errors rather than the transcription errors ASR systems are normally scored upon. Currently there has been increased interest in developing new and more sophisticated methods for determining confidence measures which make use of features other than purely acoustic ones, including confidence measures based on the posterior probability of phones generated by the decoder or estimated from N-best lists (Andorno et al., 2002), the use of word lattices (Falavigna et al., 2002) and parse-level features (Zhang and Rudnicky, 2001), the use of semantic or conceptual features (Guillevic et al., 2002; Wang and Lin, 2002) and pragmatic features to measure confidence in recognition (Ammicht et al., 2001). There has also been increased interest in the use of various machine learning techniques to combine potential features sets (Zhang and Rudnicky, 2001; Moreno et al., 2001).

In this paper we extend the set of features that can be used to predict recognition error still further. We examine the role of prosody as an indicator of both transcription and conceptual error. We focus on prosodic features primarily for several reasons. First, ASR performance is known to vary widely based upon speaking style or context of speaking (Weintraub et al., 1996), speaker gender and age, and native vs. non-native speaker status. All of these observed differences have their prosodic component, which may play a role in the deviation of the new speech produced by women, children, or non-native speakers, or spoken in a casual speaking style, from the speech data on which most ASR systems have historically been trained. Prosodic differences have been found to characterize differences between speaking styles (Bruce, 1995; Hirschberg, 1995), such as casual vs. formal speech (Blaauw, 1992), and between individual speakers (Kraayeveld, 1997). Second, as noted above, a number of studies (Wade et al., 1992; Oviatt et al., 1996; Swerts and Ostendorf, 1997; Levow, 1998; Bell and Gustafson, 1999)

report that hyperarticulated speech, characterized by careful enunciation, slowed speaking rate, and increase in pitch and loudness, often occurs when users in human–machine interactions try to correct system errors. Others have demonstrated that such speech also decreases recognition performance (Soltau and Waibel, 1998) and that compensation for it can improve performance (Soltau and Waibel, 2000; Soltau et al., 2002). Prosodic features have also been shown to be effective in ranking recognition hypotheses, as a post-processing filter to score ASR hypotheses (Hirschberg, 1991; Veilleux, 1994; Hirose, 1997). We hypothesize that misrecognitions might differ in their prosody from correctly recognized turns—perhaps due to prior misrecognitions—and thus might be identifiable in prosodic terms.

In Section 2 we describe our corpora. In Section 3 we present results comparing prosodic analyses of correctly and incorrectly recognized speaker turns in both corpora. In Section 4 we describe machine learning experiments based on the features examined in Section 3 that explore the predictive power of prosodic features alone and in combination with other automatically available information, including information currently available to ASR systems as a result of the recognition process but not currently used in making rejection decisions. Our results indicate that there are significant prosodic differences between correctly and incorrectly recognized utterances and that these differences can in fact be used, alone and in conjunction with other automatically available or easily derivable information, to predict very accurately whether an utterance has been misrecognized. Our results also indicate that humanly perceptible hyperarticulation itself cannot account for large amounts of ASR error, although features associated with hyperarticulation such as characteristics of slow speaking rate, duration, wide F0 excursion, and loudness do appear to be significantly correlated with recognition error. We also find that, while prosodic characteristics are significantly associated with ASR error, they are more effective in predicting that error in conjunction with other features of the discourse than alone. In Section 5 we discuss our conclusions and our future research.

## 2. Corpora

Our corpora consisted of recordings from two SDSs which employed different ASR systems, the experimental TOOT SDS (Litman and Pan, 1999), which provided users with train information over the phone from an online website, and the W99 SDS (Rahim et al., 1999), which provided conference registrants with information about paper submissions and registration for the Automatic Speech Recognition and Understanding (ASRU-99) workshop.

### 2.1. The TOOT corpus

The TOOT corpus was collected using an experimental SDS for the purpose of comparing differences in confirmation strategy (explicit, implicit or no confirmation provided to the user), type of initiative supported (system, user, or mixed)<sup>1</sup> and whether or not these strategies could be changed by the user during the dialogue using voice commands; for example, if a user wished to change the system strategy to one of system initiative during a dialogue, the user could say *Change strategy* followed by *System*. TOOT is implemented using a platform developed at AT&T combining ASR, text-to-speech, a phone interface, and modules for specifying a finite-state dialogue manager, and application functions (Kamm et al., 1997). The speech recognizer is a speaker-independent hidden Markov model system with context-dependent phone models for telephone speech and constrained grammars defining the vocabulary that is permitted for each dialogue state. Confidence scores for recognition were available only at the turn, not the word, level (Zeljko, 1996) and were based on acoustic likelihoods only. Thresholds were set differently for different grammar states, after some experimentation with the system. An example TOOT dialogue is shown in Fig. 1. In this version of the system, the user is allowed to

<sup>1</sup> Either the system or the user controls the course of the dialogue, such as what will be talked about next, or this control is shared.

---

TOOT: Hi, this is AT&T Amtrak schedule system. This is TOOT.  
How may I help you?  
User: I want the trains from New York City to Washington DC on  
Monday at 9:30 in the evening.  
TOOT: Do you want me to find the trains from New York City to  
Washington DC on Monday approximately at 9:30 in the  
evening now?  
User: Yes.  
TOOT: I am going to get the train schedule for you ...

---

Fig. 1. Example dialogue excerpt from TOOT.

---

...  
User: I'd like to go to Chicago from Baltimore on Saturday at 8 am.  
TOOT: Do you want to leave in the morning or in the evening?  
User: Uh morning.  
TOOT: Sorry, I can't understand you. Please repeat your answer.  
User: In the morning.  
TOOT: Do you want me to find the trains from Baltimore to Chicago on  
Saturday around 10 morning now?  
User: No.  
TOOT: How may I help you?  
...

---

Fig. 2. Example dialogue excerpt with misrecognitions.

take the initiative and the system provides no confirmation—except before the system queries the train database. Fig. 2 shows how this version of the system behaves when user utterances are both rejected and misrecognized by the system. An excerpt using another version of TOOT, in which the system takes the initiative and users are given explicit confirmation of their input, is presented in Fig. 3.

Subjects were asked to perform four tasks with one of six versions of TOOT, three combinations of confirmation type and locus of initiative (system initiative with explicit system confirmation, user

initiative with no system confirmation until the end of the task, mixed initiative with implicit system confirmation), with variants of these three that were either fixed for the duration of the task or in which the user could switch to a different confirmation/initiative strategy using voice commands. The task scenario for the dialogue shown in Fig. 2, for example, was: “Try to find a train going to **Chicago from Baltimore on Saturday at 8 o'clock am**. If you cannot find an exact match, find the one with the **closest** departure time. Please write down the **exact departure time** of the train you found as well as the **total travel time**”.

---

TOOT: Welcome to AT&T Amtrak schedule system. This is TOOT.  
Which city do you want to go to?  
User: Washington DC.  
TOOT: Do you want to go to Washington DC?  
...

---

Fig. 3. Dialogue excerpt from system initiative/explicit confirmation strategy version of TOOT.

Subjects were 39 students, 20 native speakers of standard American English and 19 non-native speakers; 16 subjects were female and 23 male. The exchanges were recorded and the behavior of both system and user was logged automatically. Mean length of subject turns was 1.92 s and 3.70 words. The corpus of user turns was 12 h long. There were a total of 152 tasks, with mean task length of 15.32 turns across all subjects. However there was a large variation between tasks (standard deviation (sdev) = 11.36). The shortest task was only 2 turns long but the longest was 95.

All dialogues were manually transcribed and system and user turns were identified by hand as beginning and ending with the system or user output. The orthographic transcriptions were compared to the ASR (one-best) recognized string to produce a **word accuracy rate** (WA) for each turn. In addition the **concept accuracy** (CA) of each turn was labeled by the experimenters by listening to the dialogue recordings while examining the system log. In our definition of CA, if the best-scoring ASR hypothesis correctly captured all the task-related information given in the user's original input (e.g., date, time, departure or arrival cities), the turn was given a CA score of 1, indicating a *semantically correct recognition*. Otherwise, the CA score reflected the percentage of correctly recognized task concepts in the turn. For example, if the user said *I want to go to Baltimore on Saturday at 10 o'clock* but the system's best hypothesis was *Go to Boston on Saturday*, the CA score for this turn would be 0.33. While WA is the traditional method of evaluating ASR success, CA does not penalize for word errors that are unimportant to overall utterance interpretation.

For the study described below, we examined 2328 user turns from 152 dialogues generated during these experiments. 202 of the 2328 turns were **rejected** by the system because its best hypothesis was below a predefined rejection threshold based on the value of the acoustic confidence score. (The TOOT confidence score thresholds were set relatively low, so that the system tended to misrecognize rather than reject utterances.) After rejections, the system asked users to repeat their last utterance.

Seventy percentage of the 2328 turns we examined were assigned a CA score of 1 by our labelers (i.e., were conceptually accurate). The mean CA score for all turns, where CA ranged from 0 to 1, was 0.73. Sixty one percentage of turns had a WA of 1 (i.e., were exact transcriptions of the spoken turn) and mean WA score over all turns was 0.60. Even though WA was low, the system's actual ability to correctly interpret user input (i.e., the CA score) was somewhat higher.

## 2.2. The W99 corpus

The W99 corpus derives from a spoken dialogue system used to support registration and information access for the ASRU-99 workshop (Rahim et al., 1999). Unlike the TOOT experimental system, this was a live system with real users. The system was implemented using an IP and computer telephony platform, and included a speech recognizer, natural language understander, dialogue manager, text-to-speech system, and application database. The system used WATSON (Sharp et al., 1997), a speaker-independent hidden Markov model ASR system, with HMMs trained using maximum likelihood estimation followed by minimum classification error training. It rejected utterances based on their ASR confidence score, which was based on a likelihood ratio distance compared to a predefined rejection threshold, similar to confidence scoring in the TOOT system. As with the TOOT platform, ASR confidence scores were available only at the turn, not the word, level. This system employed a mixed initiative dialogue strategy: the system generally gave the user the initiative (e.g., users responded to open-ended system prompts such as *What can I do for you?*), but could take the initiative back after ASR problems (e.g., giving users directed prompts such as *Please say . . .*). A sample dialogue appears in Fig. 4.

Since the initial version of W99 was built before any data collection occurred, it used acoustic models from a pre-existing call-routing application. State-dependent bigram grammars were also obtained from the same application, as well as from interactions collected using a text-only version of the system. (Subsequently, another version

---

W99: Welcome to the ASRU-99 voice automated system. What can I do for you?  
 Caller: I need to register please.  
 W99: IEEE members may register at the workshop discounted rate before September first. Are you a member of the signal processing society?  
 Caller: I uh yes I am.  
 W99: Good!, I can help you with registration. Do you have your card number available?  
 Caller: Um yep I do uh membership number is uh <card number>.  
 W99: Thank you, <Callername>! You are now registered for the ASRU-99 workshop. You will receive confirmation electronically within 2 days. Would you like to make your hotel reservation?  
 Caller: No thank you.

---

Fig. 4. Example dialogue excerpt from W99.

of the system was built, using 750 “live” utterances from the initial deployment to adapt the HMMs and grammars.) The data analyzed in this paper consist of 2997 utterances collected during an in-house trial with 50 summer students as well as the actual registration for the ASRU-99 workshop, with the recognition results generated during that process. Mean duration of user turns for this corpus (end-pointed automatically and thus less accurately than for the TOOT corpus) was 8.24 s and 4.92 words. The total length of all user utterances in the corpus was 25464.84 s (70.74 h), although, again, this included a considerable amount of silence. It was impossible to calculate mean length of task, since turns were not identified by user.

### 2.3. Comparing TOOT and W99

The W99 and TOOT corpora differ from each other in several important ways. The implementation platform and all of the major system components (ASR, TTS, dialogue management, semantic analysis) are different, with the W99 system using newer and generally more robust technology (e.g., stochastic language models instead of hand-built grammars). The TOOT data were obtained from structured experiments, while the W99 data included both experimental and non-experimental data. Finally, the W99 system used a primarily user initiative dialogue strategy with limited backoff to system initiative, while TOOT employed a wide variety of initiative and confirmation strategies.

Our descriptive analyses of the two corpora also differ in several ways, due to the nature of the data

and availability of annotation. For the TOOT corpus, we did not have access to the speech files actually sent to the recognizer during the experiments, so we end-pointed (by hand) the recordings made of both sides of the dialogue at the time, to demarcate the user turns and system turns from beginning of user or system speech to end. For the W99 data, we were able to analyze the actual speech used as input by the speech recognition system; thus our durational information was generated automatically and we did not have all of the timing information we manually annotated in the TOOT corpus full dialogue recordings as noted above. For the TOOT corpus, we had both WA and CA scores, which were not available for the W99 data; for the latter, we could only examine prosodic characteristics of recognition errors defined in terms of transcription error, not conceptual error. The W99 system however provided an automatically generated semantic accuracy score based on its assessment of its own performance, which we employed in the machine learning experiments described in Section 4. For both corpora, we examined misrecognition as a binary feature—if the recognizer made any error in its transcription or misinterpretation of a speaker turn, we counted that turn as ‘misrecognized’. A final distinction between the TOOT and W99 corpora was speaker identity information. For TOOT we could identify which speaker produced each turn, but for the W99 data such information was not collected.

Table 1 compares the two corpora overall in terms of the acoustic and prosodic features we will examine in this study. These features are defined below in Section 3. Suffice it to note here that they

Table 1  
Comparison of prosodic and acoustic features of the TOOT and W99 corpora

Feature	TOOT mean	TOOT sdev	W99 mean	W99 sdev	<i>P</i>
<b>F0 Max</b> (Hz)*	227.16	44.09	232.37	93.43	0.04
<b>F0 Mean</b> (Hz)*	162.96	76.74	145.77	43.79	0
<b>RMS Max</b> (A)*	1612.47	1020.38	2002.94	1834.11	0
<b>RMS Mean</b> (A)*	395.75	261.26	652.14	537.89	0
<b>Dur</b> (s)*	1.92	2.44	8.24	5.01	0
<b>PPau</b> (s)	0.71	0.63	NA	NA	NA
<b>Tempo</b> (sps)*	2.48	1.37	0.98	1.09	0
<b>% Silence</b> *	0.44	0.17	0.80	0.18	0

\*Significant at a 95% confidence level ( $p \leq 0.05$ ).

include a variety of features related to pitch range, timing, and perceived loudness. The comparison in Table 1 shows that, in every prosodic and acoustic feature that we could calculate for each corpus, these corpora differ significantly.<sup>2</sup> We hypothesize that, if misrecognized turns differ from correctly recognized turns in both corpora in terms of similar features, it will thus be likely that this difference is a relative and not an absolute one.

### 3. Distinguishing correct from incorrect recognitions

#### 3.1. Transcript and concept errors in the TOOT corpus

For the TOOT corpus, we looked for distinguishing prosodic characteristics of misrecognitions, defining misrecognitions in two ways—in terms of word accuracy (turns with  $WA < 1$ ) and in terms of concept accuracy (turns with  $CA < 1$ ). As noted in Section 1, previous studies have speculated that hyperarticulated speech (slower and louder speech which contains wider pitch excursions) may follow recognition failure and be associated with subsequent failures. So, we examined the following prosodic features for each user turn, which we felt might be good indicators of hyperarticulated speech:

- maximum and mean fundamental frequency values (**F0 Max**, **F0 Mean**),

- maximum and mean energy values (**RMS Max**, **RMS Mean**),
- total duration (**Dur**),
- length of pause preceding the turn (**PPau**),
- speaking rate, calculated in syllables per second (sps) (**Tempo**),
- amount of silence within the turn (**% Silence**).

F0 and RMS values, representing measures of pitch excursion and loudness, were calculated from the output of Entropic Research Laboratory's pitch tracker, *get\_f0* (Talkin, 1995),<sup>3</sup> with no post-correction. Timing variation was represented by the following features: Duration of a speaker turn (**Dur**) and length of pause between system and speaker turns (**PPau**) were computed from the temporal labels associated with each turn's beginning and ending (cf. Section 2.1). **Tempo** was approximated in terms of syllables in the recognized string per second, while **% Silence** was defined as the percentage of zero-valued F0 frames in the turn, taken from the output of the pitch-tracker, and representing roughly the percentage of time within the turn that the speaker was silent.

To ensure that our results were speaker independent, we performed within-speaker comparisons and analyzed these across speakers in the following way: We first calculated mean values for each prosodic feature for the set of recognized turns and their misrecognized turns for each individual speaker. So, for speaker A, we divided all turns produced in the four tasks into two classes,

<sup>2</sup> In the table, *P* is the likelihood that the difference between the two means for each feature is due to chance.

<sup>3</sup> *get\_f0* and other Entropic software is currently available free of charge at <http://www.speech.kth.se/esps/esps.zip>

based on whether the ASR system had correctly recognized that turn or not. For each class, we then calculated mean **F0 Max**, mean **F0 Mean**, and so on. After this step had been repeated for each speaker and for each feature, we then created two vectors of speaker means for each individual prosodic feature—e.g., a vector containing the mean **F0 Max** for each speaker's recognized turns and a corresponding vector containing the mean **F0 Max** for each speaker's misrecognized turns. We then performed paired *t*-tests on the two vectors for each feature, to see if there were similar significant differences in individual speaker's prosodic features for the two classes of turn, across all speakers. Table 2 shows results of these analyses for WA-defined recognition errors. From Table 2 we see that speaker turns containing transcription errors exhibit, on average, larger pitch excursions (**F0 Max**) and greater amplitude excursions (**RMS Max**) than those that are correctly recognized. They are also longer in duration (**Dur**), are preceded by longer pauses (**PPau**), and are spoken at a slower rate (**Tempo**). That is, they are higher in pitch, louder, longer, follow longer pauses, and are slower than turns that contain no transcription errors.

Comparing these findings with those for CA-defined misrecognition in Table 3, we see a similar picture. These misrecognitions also differ significantly from correctly recognized turns in terms of the same prosodic features as those in Table 2 (**F0**

Table 2  
Comparison of misrecognized (WA<1) vs. recognized turns by prosodic feature across speakers

Feature	<i>T</i> -stat	Mean mis-recognized–recognized	<i>P</i>
<b>F0 Max*</b>	5.78	25.84 Hz	0
<b>F0 Mean</b>	1.52	1.56 Hz	0.14
<b>RMS Max*</b>	2.51	150.56	0.02
<b>RMS Mean</b>	−1.82	−25.05	0.08
<b>Dur*</b>	9.94	2.13 s	0
<b>PPau*</b>	5.86	0.29 s	0
<b>Tempo*</b>	−4.71	−0.54 sps	0
<b>% Silence</b>	−1.48	−0.02%	0.15

df = 38 in each analysis

\*Significant at a 95% confidence level ( $p \leq 0.05$ ).

Table 3  
Comparison of misrecognized (CA < 41) vs. recognized turns by prosodic feature across speakers

Feature	<i>T</i> -stat	Mean mis-recognized–recognized	<i>P</i>
<b>F0 Max*</b>	5.39	27.84 Hz	0
<b>F0 Mean</b>	1.76	2.01 Hz	0.09
<b>RMS Max*</b>	2.68	167.17	0.01
<b>RMS Mean</b>	−1.58	−24.18	0.12
<b>Dur*</b>	9.21	2.10 s	0
<b>PPau*</b>	5.79	0.35 s	0
<b>Tempo*</b>	−4.36	−0.54 sps	0
<b>% Silence</b>	−1.30	−0.02%	0.20

df = 37 in each analysis

\*Significant at a 95% confidence level ( $p \leq 0.05$ ).

**Max**, **RMS Max**, **Dur**, **PPau**, and **Tempo**). So, whether defined by WA or CA, misrecognized turns exhibit significantly higher F0 and RMS maxima, longer durations, longer preceding pauses, and slower rates than correctly recognized speaker turns. The common features which distinguish both types of misrecognition are consistent with the hypothesis that there is a strong association between misrecognition and hyper-articulation.

While the comparisons in Tables 2 and 3 were made on the means of raw values for all prosodic features (**Raw**), little difference is found when values are normalized by dividing by the value of (the same speaker's) first or preceding turn in the dialogue.<sup>4</sup> For these analyses it appears to be the case that relative differences in speakers' prosodic values, not deviation from some 'acceptable' range, distinguishes recognition failures from successful recognitions. A given speaker's turns that are higher in pitch or loudness, or that are longer, or that follow longer pauses, are less likely to be

<sup>4</sup> The only differences in the comparison of normalized features to raw features occur for WA-defined misrecognition, where **Tempo** is not significantly different when features are normalized by preceding turn and in CA-defined misrecognition, where preceding pause is not significantly different when normalizing by the first turn in the task and **Tempo** is not different when normalizing by preceding turn.

recognized correctly than that same speaker's turns that are lower in pitch or loudness, shorter, and follow shorter pauses—however correct recognition is defined.

We further explored the hypothesis that hyperarticulation leads to misrecognition, since the features we found to be significant indicators of failed recognitions (F0 excursion, loudness, long preceding pause, longer duration, and tempo) are all features previously associated with hyperarticulated speech. Recall that earlier work has suggested that speakers may respond to failed recognition attempts by hyperarticulating, which itself may lead to more recognition failures. Had our analyses simply identified a means of characterizing and identifying hyperarticulated speech in terms of its distinguishing prosodic features? What we found suggested a more complicated picture of the role of hyperarticulated speech in recognition errors.

Before performing our acoustic analyses, we had independently labeled all speaker turns for evidence of hyperarticulation. Two of the authors labeled each turn as “not hyperarticulated”, “some hyperarticulation in the turn”, and “hyperarticulated”, using the criteria and methods of (Wade et al., 1992), without reference to information about recognition error or prosodic features. 24.1% of the turns in our corpus exhibit some indication of hyperarticulation (i.e., were labeled by at least one labeler as showing some hyperarticulation). Indeed, our data show that hyperarticulated turns are misrecognized more often than non-hyperarticulated turns (59.5% vs. 32.8%, for WA-defined misrecognition and 50.7% vs. 24.1% for CA-defined misrecognition). However, this does not in itself explain our overall results distinguishing misrecognitions from correctly recognized turns.

We replicated the preceding analyses, excluding any turn either labeler had labeled as partially or fully hyperarticulated, again performing paired *t*-tests on mean values of misrecognized vs. recognized turns for each speaker. We discovered that, in fact, for both WA-defined and CA-defined misrecognitions, when hyperarticulated turns are excluded from the analysis, essentially the same significant differences are found between correctly

and incorrectly recognized speaker turns.<sup>5</sup> Our findings for the prosodic characteristics of recognized and of misrecognized turns thus hold even when perceptibly hyperarticulated turns are excluded from the corpus. We hypothesize that hyperarticulatory trends not identifiable as such by human labelers may in fact be important for machine recognition here, and that human thresholds for perceived hyperarticulation differ from machine thresholds when test data exceeds the bounds of the training set in terms of pitch excursion, loudness, duration or tempo.

### 3.2. Transcript errors in the W99 corpus

As with the TOOT corpus, we examined the W99 corpus to see whether prosodic features distinguished misrecognitions from correctly recognized utterances. While for the TOOT corpus we were able to define recognition error in terms of both transcription and concept accuracy, for the W99 corpus we had only WA scores. We thus present results below only for WA-defined misrecognitions and compare these only to the WA-defined case for our TOOT data. We also did not have speaker identification for W99 turns. We thus could not identify the set of all turns for a particular speaker, and are not able to follow the procedure described in Section 3.1 to ensure the speaker-independence of our analysis. Instead we had to collapse data from all speakers into a single pool. Normalization of features by first turn in task could not be performed, and normalization by preceding turn introduces some noise into the data, since the preceding turn may in fact be that of a different speaker. However, for the W99 corpus we did have access to the speech files actually segmented by the system. Again, our unit of analysis was the speaker turn, but this time the speech included in the turn was defined by what the recognizer recorded and transcribed. So, for

<sup>5</sup> For WA-defined misrecognition, **RMS Mean** is also significantly different, but exactly the same features distinguish CA-defined misrecognitions from correct recognitions when hyperarticulated turns are removed, as when they are included.

Table 4  
Differences between prosodic features of misrecognized (WA < 1) vs. recognized turns for W99 corpus

Feature	<i>T</i> -stat	Mean mis-recognized–recognized	<i>P</i>
<b>F0 Max*</b>	7.07	23.81 Hz	0
<b>F0 Mean</b>	0.06	−0.10 Hz	0.95
<b>RMS Max*</b>	4.90	335.48	0
<b>RMS Mean</b>	0.34	6.70	0.7352
<b>Dur*</b>	10.55	1.88 s	0
<b>PPau</b>	NA	NA	NA
<b>Tempo*</b>	7.23	0.28 sps	0
<b>% Silence*</b>	9.58	−0.06%	0

df = 3087 in each analysis

\*Difference significant at a 95% confidence level ( $p \leq 0.05$ ).

the W99 data, we were able to use exactly what the ASR engine used for recognition in our analysis.

For each speaker turn we examined the same prosodic features we had examined for the TOOT corpus, except for **PPau**:<sup>6</sup> maximum and mean fundamental frequency values (**F0 Max**, **F0 Mean**); maximum and mean energy values (**RMS Max**, **RMS Mean**); total turn duration (**Dur**); speaking rate (**Tempo**); and amount of silence within the turn (**% Silence**). The definition of and method for calculating each of these features was that described in Section 3.1.

Since the W99 data did not contain explicit speaker identification for a given session, we collapsed all data from all sessions into a single pool, divided that pool into correct and incorrect recognitions, and performed *t*-tests on the means for each prosodic feature. Results were very similar to our analysis of the TOOT data, where we were able to calculate means for each feature on a per speaker basis. Table 4 presents prosodic differences between correct and incorrect recognitions for the W99 corpus. Comparing these results with those for the TOOT corpus in Table 2, we find few differences between the two—despite the considerable differences in the way data points were calculated, and the major differences between the two systems from which the data were obtained.

<sup>6</sup> As noted earlier, we did not have access to full recordings of the dialogue, only to the user's machine-endpointed speech.

The amount of turn-internal silence (**% Silence**) distinguishes misrecognized turns in the W99 data, though not in the TOOT corpus. While misrecognized turns are slower than correctly recognized turns in the TOOT corpus—consistent with hyper-articulation—the opposite is true in the W99 corpus. Even if we calculate tempo for the TOOT data in the same way as for the W99 data (merging **Dur** and **PPau**), misrecognized turns are still significantly slower than correctly recognized turns ( $T$ -stat = 3.21,  $p \leq 0.003$ ), although the difference between the means is halved (−0.26 sps). Table 1 showed that there are significant differences in the two corpora in tempo, as in all other prosodic and acoustic features compared. We hypothesize some significant difference in the mismatch between the training data of the recognizers for the two corpora and the actual data of the corpora themselves; that is, that the recognizer used for TOOT recognized faster speech better than slower, but the opposite was true for the recognizer used for W99. In all other respects that we can measure, the two corpora lead us to the same conclusions about the relationships among prosodic features and misrecognized speaker turns: that pitch excursions are more extreme (higher F0 maximum) for misrecognized turns than for correctly recognized turns, misrecognized turns contain louder portions (higher RMS maximum), and misrecognized turns are longer.

The results presented in Table 4 are based on means for raw values of prosodic features in each turn. Recall from Section 3.1 that, for the TOOT corpus, we found little difference between using raw scores and scores normalized by value of first or of preceding turn. For the W99 corpus this picture is somewhat different. While means calculated on the absolute values for **Dur**, **RMS Max**, **F0 Max**, **Tempo**, and **% Silence** distinguish misrecognitions from recognitions, when these values are normalized by preceding turn, only **Dur**, **F0 Max** and **Tempo** significantly distinguish the two groups of turns. Since there was some noise in the data due to lack of identifiable boundaries for each speaker's interaction, this difference may not be too reliable. For the W99 data, we are also pooling data from all speakers rather than identifying within-speaker differences and generalizing over

---

Prosodic (PROS) :
Raw (raw values): F0 Max, F0 Mean, RMS Max, RMS Mean, Dur, PPau, Tempo, % Silence
Norm1 (values normalized by first turn in dialogue): F0 Max1, F0 Mean1, RMS Max1, RMS Mean1, Dur1, PPau1, Tempo1, % Silence1
Norm2 (values normalized by previous turn in dialogue): F0 Max2, F0 Mean2, RMS Max2, RMS Mean2, Dur2, PPau2, Tempo2, % Silence2
Position (POS) : turn
Speech Recognition (ASR) :
Original : grammar, string, confidence
StrDerived : yn, no, cancel, help, words, syls
System Settings (SYS) : initiative, confirmation, adaptability, strategies
Manually Labeled (MAN) : hyperarticulation, task, subject, gender, native

---

Fig. 5. Feature set for predicting misrecognitions.

these. On the whole, we suspect that the TOOT results may be more reliable.

#### 4. Predicting misrecognitions

Given the prosodic differences between misrecognized and correctly recognized utterances in our corpora, is it possible to predict accurately when a particular utterance will be misrecognized or not? This section describes experiments using the machine learning program *RIPPER* (Cohen, 1996) to automatically induce prediction models, using prosodic as well as additional features. Like many learning programs, *RIPPER* takes as input the classes to be learned, a set of feature names and possible values (symbolic, continuous, or text), and training data specifying the class and feature values for each training example. *RIPPER* outputs a classification model for predicting the class of future examples. The model is learned using a greedy search procedure, guided by an information gain metric, and is expressed as an ordered set of if-then rules.

##### 4.1. Predicting transcript and concept errors in the TOOT corpus

For our machine learning experiments, the predicted classes correspond to correct recognition (T) or not (F). As in Section 3, we examine both WA-defined and CA-defined notions of correct

recognition for the TOOT corpus. We also represent each user turn as a set of features. We first describe the set of features for the TOOT machine learning experiments, followed by our results for predicting transcript (WA-defined) and concept (CA-defined) errors, respectively.

##### 4.1.1. Features

The entire feature set used in our learning experiments is presented in Fig. 5. The feature set includes the automatically computable raw and normalized versions of the prosodic features in Tables 2 and 3 (which we will refer to as **PROS**), and a manually computed feature representing the sum of the two scores from the hyperarticulation labeling discussed in Section 3 (the feature **hyperarticulation**).<sup>7</sup> The feature set also includes several other types of non-prosodic potential predictors of misrecognition. The feature **turn** represents the distance of the current turn from the beginning of the dialogue, while a number of ASR features are derived from standard inputs and outputs of the speech recognition process. They include **grammar**, the identity of the finite-state grammar used as the ASR language model for the dialogue state the system expected the user to be in (e.g., after the

<sup>7</sup> Not all of the features we term “automatically computable” were in fact automatically computed in our data, due to the necessity of manually end-pointing the TOOT user turns, as discussed in Section 3.

system produced a *yes–no* question, the user's next turn would be recognized with the *yes–no* grammar), the string the recognizer proposed as its best hypothesis (**string**), and the associated turn-level acoustic confidence score produced by the recognizer (**confidence**). We included these features as a baseline against which to test new methods of predicting misrecognitions, although, currently, we know of no ASR system that includes the identity of the recognized string in its rejection calculations.<sup>8</sup> As subcases of the **string** feature, we derive features representing whether or not the recognized string includes variants of *yes* or *no* (**yn**), any variant of *no* such as *nope* (**no**), and the special dialogue management commands *cancel* (**cancel**) and *help* (**help**). We also derive features to approximate the length of the user turn in words (**words**) and in syllables (**sylls**) from the **string** feature. Both are positively correlated with turn duration (**words**:  $t = 35.33$ ,  $df = 2253$ ,  $p = 0$ ,  $r = 0.60$ ; and **sylls**:  $t = 36.44$ ,  $df = 2253$ ,  $p = 0$ ,  $r = 0.61$ ). Finally, we include a set of features representing the system's dialogue manager settings when each turn was collected (**SYS**). These features include the system's current initiative and confirmation strategies (**initiative**, **confirmation**), whether users could adapt the system's dialogue strategies, as described in Section 2.1 (**adaptability**), and the combined initiative/confirmation strategy in effect at the time of the utterance (**strategies**). A set of experimental features that would not be automatically available in a non-experimental setting are also considered, namely the task the user was performing (**task**) and some user-specific characteristics, including the subject's identity and gender (**subject**, **gender**), and whether or not the subject was a native speaker of American English (**native**). We included these features to determine the extent to which particulars of task, subject, or interaction style influenced ASR success rates or our ability to predict them; previous work showed that some of these factors affected

<sup>8</sup> While the entire recognized string is provided to the learning algorithm, **RIPPER** rules test for the presence of particular words in the string.

TOOT's performance (Litman and Pan, 1999; Hirschberg et al., 1999).

#### 4.1.2. Predicting transcript errors

Table 5 shows the relative performance of a number of the feature sets we examined, and for comparison also presents two relevant baselines; results here are for misrecognition defined in terms of WA. The table shows a mean error rate and the associated standard error of the mean (SE)<sup>9</sup> for the classification model that was learned from each feature set; these figures are based on 25-fold cross-validation.<sup>10</sup>

The simplest baseline classifier for misrecognition, predicting that the recognizer is always correct (the majority class of T), has a classification error of 39.22%. Since TOOT itself used the features **grammar** and **confidence** to predict misrecognitions, TOOT's actual performance during the experiment provides a more realistic baseline. Whenever the confidence score fell below a grammar-specific threshold (manually specified by the system designer), TOOT asked the user to repeat the utterance. Analyzing these rejected utterances shows that TOOT incorrectly rejected 17 correct recognitions, and did not reject 736 misrecognitions—a total error rate in classifying misrecognitions of 32.35%. We term this the TOOT baseline.

The best performing feature set includes only the raw prosodic and ASR features and reduces the TOOT baseline error to an impressive 8.64%.

<sup>9</sup> The standard error of the mean is the standard deviation of the sampling distribution (the different sample estimates) of the mean. It can be estimated from a single sample of observations as the standard deviation of the observations divided by the square root of the sample size.

<sup>10</sup> In 25-fold cross-validation, the total set of examples in our corpus is randomly divided into 25 disjoint test sets, and 25 runs of the learning program are performed. Thus, each run uses the examples not in the test set for training and the remaining examples for testing. An estimated error rate is obtained by averaging the error rate on the testing portion of the data from each of the 25 runs. Ninety five percentage confidence intervals are then approximated for each mean, using the mean plus or minus twice the standard error of the mean. When two errors plus or minus twice the standard error do not overlap, they are statistically significantly different.

Table 5  
Estimated error for predicting misrecognized turns (WA < 1)

Features used	Error (%)	SE
<b>Raw + ASR</b>	8.64	0.53
<b>ALL-(Dur, syls, words)</b>	9.67	0.62
<b>ALL</b>	9.88	0.66
<b>Raw + string + StrDerived</b>	11.98	0.55
<b>Raw + grammar + confidence</b>	12.33	0.68
<b>Raw + confidence</b>	12.76	0.71
<b>ASR</b>	14.83	0.81
<b>String + StrDerived</b>	18.00	0.86
<b>Grammar + confidence</b>	18.70	1.03
<b>Confidence</b>	18.91	1.00
<b>Raw</b>	19.20	0.80
<b>PROS</b>	19.98	0.86
<b>Dur</b>	20.92	0.85
<b>Tempo</b>	26.93	0.80
<b>Norm2</b>	28.57	1.08
<b>Norm1</b>	32.17	0.81
<b>Native</b>	39.22	1.07
TOOT baseline	32.35	
Majority baseline	39.22	

The use of this learned rule-set could have yielded an extremely dramatic improvement in TOOT's performance. The performance of this feature set is not improved by identifying individual subjects or their characteristics, such as **gender** or native/non-native status (**native**), or by adding other manually labeled features such as **hyperarticulation**, or by distinguishing among system dialogue strategies: the feature set corresponding to all features (**ALL**) yielded the statistically equivalent 9.88%. The estimated error for the raw prosodic and **ASR** features is significantly lower than the estimated error for all of the remaining feature sets (below **ALL** in the table).

Examining some of these remaining feature sets, Table 5 shows that using raw prosodic features in conjunction with **ASR** features (error of 8.64%) significantly outperforms the set of raw prosodic features alone (error of 19.2%), which in turn outperforms (although not significantly) any single prosodic feature. **Dur** is the best such feature, with an error of 20.92%, and significantly outperforms the second most useful single feature, **Tempo**, which has an estimated error of 26.93%. The importance of duration as a signal of ASR error is,

of course, not surprising in itself, since it correlates with turn length in words, and longer utterances have a greater chance of being misrecognized in terms of WA. Even when **Dur** itself and all correlated features, such as length in words and syllables, are removed from the feature set, performance degrades only marginally, to 9.67% estimated error; this performance is second only to the best feature set, combining all raw prosodic and **ASR** features. The rest of the single prosodic features, as well as the **hyperarticulation** feature, yield errors below the actual TOOT baseline, and thus are not included in the table. (In these machine learning experiments we are unable to compare prosodic features' effect on ASR performance on a per speaker basis, which is where our descriptive statistical analyses found significant differences in prosodic features other than duration.) The native/non-native distinction (**native**), which often affects recognition performance, is also not useful as a predictor of recognition error here, performing about as well as the majority baseline classifier.

The **unnormalized** raw prosodic features significantly outperform the normalized versions by 9–13%. Recall that prosodic features normalized by first utterance in task and by previous utterance showed little performance difference in the analyses described in Section 3. This difference may indicate that, for a given recognizer, there are indeed limits on the ranges in features such as **F0 Max** and **RMS Max**, **Dur** and **PPau** within which recognition performance is optimal, defined by the recognizer's training data. It seems reasonable that extreme deviation from characteristics of the acoustic training material should in fact impact ASR performance, and our experiments may have uncovered, if not the critical variants, at least important acoustic correlates of them. Finally, using the raw prosodic features is almost identical to simultaneously using all three forms of the prosodic features (**PROS**).

A comparison of other rows in our table can help us to understand what prosodic features are contributing to misrecognition identification, relative to the more traditional ASR techniques. Do our prosodic features simply correlate with information already in use by ASR systems

---

```

if (confidence ≤ -2.85) ∧ (Dur ≥ 1.27) then F
if (confidence ≤ -4.34) then F
if (Tempo ≤ .81) then F
if (confidence ≤ -4.09) then F
if (confidence ≤ -2.46 ∧ (string contains 'help')) then F
if (confidence ≤ -2.47 ∧ (PPau ≥ .77) ∧ (Tempo ≤ .25)) then F
if (string contains 'nope') then F
if (Dur ≥ 1.71) ∧ (Tempo ≤ 1.76) then F
else T

```

---

Fig. 6. Rule-set for predicting misrecognized turns (WA < 1) from raw prosodic and ASR features.

(e.g., confidence score, grammar), or at least available to them (e.g., recognized string)? First, the error using ASR confidence score alone (18.91%) is significantly worse than the error when prosodic features are combined with ASR confidence scores (12.76%)—and is also comparable to the use of prosodic features alone (19.2%). Similarly, using ASR confidence scores and **grammar** (18.70%) is comparable to using prosodic features alone (19.2%), but significantly worse than using **confidence**, **grammar**, and prosody (12.33%).<sup>11</sup> Thus, prosodic features in conjunction with traditional ASR features significantly outperform these traditional features alone for predicting WA-based misrecognitions. When used alone, the prosodic features perform comparably to the traditional features.

Another interesting finding from our table is the predictive power of information available to current ASR systems but not made use of in calculating rejection likelihoods—the identity of the recognized string. It seems that, at least in our task and for our ASR system, the appearance of certain particular recognized strings is an extremely useful cue to recognition accuracy. Using our string-based features in conjunction with the traditional ASR features (error of 14.83%) significantly outperforms using only the traditional ASR features (error of 18.70%). Even using only the **string** and its derived features (error of 18%) outperforms

using **grammar** and **confidence** (error of 18.70%) (although not with statistical significance). So, even by making use of information currently available from the traditional ASR process, ASR systems could improve their performance on identifying rejections by a considerable margin. A caveat here is that the string-based features, like grammar state, are unlikely to generalize from task to task or recognizer to recognizer, but these findings suggest that such features should be considered as a means of improving rejection performance in stable systems.

The classification model learned from the best performing feature set in Table 5 is shown in Fig. 6.<sup>12</sup> The first rule **RIPPER** finds with this feature set is that if the acoustic confidence score is less than or equal to  $-2.85$ , and if the user turn is at least 1.27 s, then predict that the turn will be misrecognized.<sup>13</sup> As another example, the seventh rule says that if the **string** contains the word *nope* (and possibly other words as well), also predict misrecognition. While three prosodic features appear in at least one rule (**Dur**, **Tempo**, and **PPau**), the features shown to be significant in our statistical analyses (Section 3) are not the same features as in the rules. As noted above, it is difficult to compare our machine learning results with the statistical analyses, since (a) the statistical analyses looked at only a single prosodic variable at a time, and (b) data points for that analysis were means

<sup>11</sup> Recall that TOOT predicted misrecognitions using only **confidence** and **grammar**. The fact that TOOT's baseline error rate was 32.35% suggests that the manual specification of grammar-dependent confidence thresholds could have been greatly improved using machine learning (18.70%).

<sup>12</sup> Rules are presented in order of importance in classifying data. When multiple rules are applicable, **RIPPER** uses the first rule.

<sup>13</sup> The confidence scores observed in our data ranged from a high of  $-0.09$  to a low of  $-9.88$ .

calculated per speaker, while the learning algorithm operated on all utterances, allowing for unequal contributions by speaker. Besides the three prosodic features, the rule-set also contains two of the ASR features (**confidence** and **string**). While the rules are not specific to the particular conditions of, and the participants in, the corpus collection, the fifth and seventh rules use the feature **string**, which as discussed above might make them less likely to generalize.

The classification model learned from the feature set **Raw** in Table 5 is shown in Fig. 7. When only the prosodic features are used for misrecognition prediction, the same three prosodic features noted above occur in the rules (**Dur**, **Tempo**, **PPau**), as does % **Silence**.

#### 4.1.3. Predicting concept errors

Results of our learning experiments with misrecognition defined in terms of CA rather than WA show the overall role of the features which predict WA-defined misrecognition to be somewhat less successful in predicting CA-defined error. Table 6 shows the relative performance of the same feature sets discussed above, with misrecognition now defined in terms of  $CA < 1$ . Here, the TOOT baseline was 24.66%: TOOT incorrectly rejected 29 correct recognitions, and did not reject 545 misrecognitions. As with the WA experiments, the best performing feature set makes use of prosodic and ASR-derived features. However, the predictive power of prosodic features combined with standard ASR features decreases when misrecognition is defined in terms of CA, to 10.48% estimated error for raw prosodic features plus ASR features. The error rate using ASR confidence alone (15.12%) is now significantly better than the error obtained using prosody alone

Table 6  
Estimated error for predicting misrecognized turns ( $CA < 1$ )

Features used	Error (%)	SE
<b>Raw + ASR</b>	10.48	0.57
<b>ALL-(Dur, syls, words)</b>	10.70	0.72
<b>ALL</b>	10.95	0.63
<b>Raw + grammar + confidence</b>	11.26	0.69
<b>Raw + confidence</b>	11.86	0.69
<b>ASR</b>	12.88	0.65
<b>Grammar + confidence</b>	14.90	0.78
<b>Confidence</b>	15.12	0.85
<b>Raw + string + StrDerived</b>	15.33	0.73
<b>String + StrDerived</b>	15.98	0.65
<b>PROS</b>	18.99	0.83
<b>Raw</b>	18.99	0.84
<b>Dur</b>	20.74	0.90
<b>Tempo</b>	25.47	0.67
<b>Norm2</b>	25.69	1.02
<b>Norm1</b>	28.09	0.90
<b>Native</b>	30.49	0.99
TOOT baseline	24.66	
Majority baseline	30.50	

(18.99%), while before they were comparable. This is particularly interesting since ASR confidence scores are intended to predict WA rather than CA. The string-based features also appear to be of less utility in this context. On the other hand, adding the prosodic features to the traditional methods still significantly improves performance (adding prosody to ASR **confidence** reduces the error from 15.12% to 11.86%, while adding prosody to **confidence** and **grammar** reduces the error from 14.9% to 11.26%). When we remove all features correlated with turn length (**All-(Dur,syls,words)**), we obtain the second highest performance of any of our feature sets, showing again that durational information, while useful, is not the only useful prosodic contribution to predicting CA. Finally,

---

```

if (Dur ≥ 1.71) ∧ (Tempo ≤ 2.00) then F
if (Dur ≥ 2.85) then F
if (Dur ≥ 1.10) ∧ (Tempo ≤ 2.34) ∧ (% Silence ≤ .65) then F
if (PPau ≥ .53) ∧ (Dur ≥ 1.53) ∧ (Tempo ≥ 3.03) ∧ (Tempo ≤ 4.21) then F
if (Tempo ≥ 5.12) ∧ (Dur ≤ .76) then F
if (Dur ≤ .83) ∧ (Tempo ≤ 1.23) then F
if (Dur ≥ 1.71) ∧ (Tempo ≥ 3.26) then F
else T

```

---

Fig. 7. Rule-set for predicting misrecognized turns ( $WA < 1$ ) from raw prosodic features.

the error rate of all of the ASR features combined (12.88%) is reduced to the lowest error rate in our table when prosodic features are added (10.48%), although this is (just) not significantly different. Thus, for CA-defined misrecognitions, our experiments have uncovered only minor improvements over non-prosodic ASR rejection calculation procedures.

#### 4.1.4. Summing up

Our results from a set of machine learning experiments illustrate the utility of considering prosodic features when predicting two types of misrecognition errors in the TOOT corpus. When predicting CA-defined misrecognitions (concept errors), the use of prosody in conjunction with other features yields minor improvements in prediction accuracy. When predicting WA-defined misrecognitions (transcript errors), prosodic features alone can predict misrecognitions with an accuracy rate comparable to more traditional features. When combined with information currently available from ASR systems, performance is greatly improved via the use of prosodic features. In our TOOT data, the standard use of ASR confidence scores predicted WA-defined misrecognitions with an error rate of 18.91%. Adding new ASR features reduced the error to 14.83%, while adding prosodic features further reduced the error to 8.64%.

#### 4.2. Predicting transcript errors in the W99 corpus

We now replicate the machine learning experiments described above, in the context of the W99 corpus. Here we are interested in examining how the W99 rules compare to the rules learned for TOOT. As in the TOOT experiments, our predicted classes correspond to correct recognition (T) or not (F), and each user turn is represented as a set of features. For W99, however, misrecognition is only defined in terms of WA, and the feature set differs somewhat. The feature set for W99 consists of the following features:

- the raw (**Raw**) prosodic features and those normalized by preceding turn (**Norm2**) (cf. Section 3.2)

Table 7

Estimated error for predicting misrecognized turns (WA < 1)

Features used	Error (%)	SE
<b>ASRW99 + prompt</b>	22.77	0.59
<b>Raw + ASRW99 + prompt</b>	23.66	0.80
<b>Raw + string</b>	23.70	0.63
<b>Grammar + confidence + string</b>	23.77	0.87
<b>ALL</b>	23.91	0.85
<b>Raw + grammar + confidence</b>	24.07	0.83
<b>Raw + grammar + confidence + string</b>	24.19	0.94
<b>Raw + confidence</b>	24.35	0.87
<b>Grammar + confidence</b>	25.68	0.78
<b>Confidence</b>	26.14	0.80
<b>Raw</b>	26.17	0.73
<b>% Silence</b>	31.30	0.93
<b>Tempo</b>	31.58	0.92
<b>String</b>	32.94	0.91
<b>Norm2</b>	36.31	0.79
<b>Dur</b>	38.51	0.80
Majority baseline	39.67	

- the three original **ASR** features from TOOT (**grammar**, **string**, **confidence**) plus a new feature representing the normalized likelihood score from the decoder (**likelihood**) (**ASRW99**)
- a single system feature representing the W99 prompt that resulted in the user's turn (**prompt**)
- whether the speaker was male or female, which was manually labeled during the corpus transcription process (**gender**)

W99 varied its 43 prompts and 10 grammars according to a mixed initiative dialogue strategy; previous work has shown that system initiative can sometimes reduce a system's misrecognition rate, as it can help keep user utterances within a system's vocabulary and grammar (Walker et al., 1998).

Table 7 shows the relative performance of a number of the feature sets we examined, and also compares these results with a baseline classifier that predicts that ASR is always wrong (the majority class of F).<sup>14</sup> The first interesting result from our experiment is that our best performing feature set (**ASRW99 + prompt**, error of 22.77%)

<sup>14</sup> The errors and standard errors (SE) result from 25-fold cross-validation.

significantly outperforms the use of grammar-dependent ASR confidence scores to determine misrecognitions (**grammar + confidence**, error of 25.68%). The best performing feature set includes the system prompt that generated the user's utterance as well as other features representing the ASR processing, and yields an error that is statistically equivalent to the error of the next 7 feature sets (i.e., through **Raw + confidence**).

With respect to the performance of our prosodic features, Table 7 shows that using multiple prosodic features (**Raw**, error of 26.17%) significantly outperforms using any single prosodic feature. % **Silence** (error of 31.30%) is the best single prosodic feature, followed by **Tempo** (error of 31.58%). In isolation, the rest of the prosodic features perform no better than the majority class baseline. It is significant that **Dur**, which, for this corpus, was calculated by automatic endpointing, performs barely above the baseline, while the more accurate hand segmentation performed for the TOOT corpus did provide useful prosodic information for distinguishing recognition errors. This suggests that when accurate timing information is not available, other automatically calculated prosodic features can provide the necessary discriminatory information.

While using prosodic features in conjunction with non-prosodic features (e.g., **Raw + string**, error of 23.7%) seems to outperform the use of prosodic features alone (error of 26.17%), none of these improvements is statistically significant. Finally, the **unnormalized** prosodic features (**Raw**) significantly outperform the normalized versions (**Norm2**) by 10%.

We also are concerned with analyzing our results to provide answers to several questions examined with respect to our TOOT analysis. Does prosody provide an alternative source of information to that already used by ASR systems? Can prosodic and ASR features be combined to improve performance? First, the error using ASR confidence score alone (26.14%) and the error using prosody alone (26.17%) are comparable. Furthermore, both errors are reduced when prosodic features are combined with ASR confidence scores (24.35%). Similarly, the error using ASR confidence scores and **grammar** (25.68%) and the

error using prosody (26.17%) are reduced when the two feature sets are combined (24.07%). Thus, as with TOOT, prosodic features perform comparably to traditional ASR features, and using both types of features seems to be better than using either type in isolation (although these latter results are not quite significant at the 95% confidence level).

Another TOOT finding replicated here is the predictive power of prosody in conjunction with information available to current ASR systems but not typically made use of when determining rejections. While ASR **string** alone has an error of 32.94%, using prosody in conjunction with **string** significantly reduces the error to 23.70%. This is statistically equivalent to the best performing feature set, in contrast to the 25.68% error using the traditional ASR confidence scores and grammars. A caveat noted earlier is that the **string** feature, like **grammar**, is less likely to generalize from task to task, but even the ASR confidence score and rejection threshold will not generalize from recognizer to recognizer.

The **RIPPER** rules learned from the best performing feature set in Table 7 are shown in Fig. 8. The first rule says that if the turn is recognized by ASR as including the **string** *yes* with an acoustic confidence score  $\geq 910$ ,<sup>15</sup> then predict a correct recognition.<sup>16</sup> Of the traditional ASR features, **confidence** appears in all rules while **grammar** only appears in one rule. With respect to the new non-prosodic features that we proposed, **string** appears in all but one rule, while **prompt** appears in one rule.

The classifier learned from the feature set **Raw** in Table 7 is shown in Fig. 9. The rules contain all of the prosodic features considered in our experiment, except for **RMS Mean** (and **PPau**, which is not applicable for the W99 corpus). % **Silence** and **Tempo**, which were the best performing prosodic features in isolation, appear in seven and three

<sup>15</sup> The confidence scores observed in our data ranged from a high of 990 to a low of 280.

<sup>16</sup> The conditions of the first rule are redundant with those of the third rule. An auxiliary software package is available for **RIPPER**, which includes a post-processing program to eliminate redundant conditions in a ruleset, if desired.

---

```

if (confidence ≥ 910) ∧ (string contains yes) then T
if (confidence ≥ 860) ∧ (string contains no) then T
if (confidence ≥ 890) ∧ (string contains yes) then T
if (confidence ≥ 860) ∧ (grammar = help) then T
if (confidence ≥ 880) ∧ (string contains zero) then T
if (confidence ≥ 860) ∧ (string contains goodbye) then T
if (confidence ≥ 860) ∧ (string contains transportation) then T
if (confidence ≥ 860) ∧ (string contains registration) ∧ (prompt = do-not-understand) then T
if (confidence ≥ 860) ∧ (string contains three) then T
if (confidence ≥ 850) ∧ (string contains registration) then T
if (confidence ≥ 880) ∧ (string contains sure) then T
if (confidence ≤ 390) ∧ (string contains no) then T
else F

```

---

Fig. 8. Best performing rule-set for predicting correctly recognized turns.

---

```

if (% Silence ≥ .95) ∧ (Dur ≤ 11.82) then T
if (% Silence ≥ .97) ∧ (Tempo ≤ .34) then T
if (% Silence ≥ .90) ∧ (Dur ≤ 6.54) ∧ (RMS Max ≥ 797.31) then T
if (% Tempo ≤ .33) ∧ (Dur ≤ 6.33) then T
if (% Silence ≥ .95) ∧ (Dur ≤ 13.52) then T
if (Dur ≤ .97) then T
if (% Silence ≥ .85) ∧ (Dur ≤ 4.01) then T
if (% Tempo ≤ .20) ∧ (F0 Mean ≤ 112.285) then T
if (% Silence ≥ .56) ∧ (Dur ≤ 2.34) ∧ (RMS Max ≥ 1385.04) then T
if (% Silence ≥ .91) ∧ (Dur ≤ 11.49) ∧ (236.01 ≤ F0 Max ≤ 352.36) then T
else F

```

---

Fig. 9. Prosodic rule-set for predicting correctly recognized turns.

rules, respectively. Duration, which was no better than the baseline in isolation, appears in eight rules. **Dur** thus appears to be a useful predictive feature in conjunction with other prosodic features. All of the features shown to be significant in our statistical analysis (Section 3) occur in the rules. The feature **F0 Mean** also occurs in one rule, even though it was not significant in Section 3.

It is also interesting to compare the W99 results with our previous results on TOOT, which used older ASR technology and poorer off-the-shelf acoustic and language models. Several of our results generalize across the TOOT and W99 experiments: the use of prosody and ASR confidence is better for predicting misrecognitions than using **confidence** alone, and the use of multiple prosodic features outperforms any single prosodic feature. In TOOT, however, the most predictive single prosodic feature was **Dur** (rather than % **Silence** as in W99), and the best performing feature set overall included prosodic features. This sup-

ports a hypothesis that the utility of the prosodic features is inversely related to the quality of the ASR models. Finally, while the specific rules we have derived for TOOT and W99 may be domain-specific in their particulars, we have demonstrated that very similar feature sets can be used to derive rules that greatly improve recognition accuracy in both domains.

## 5. Discussion

A statistical comparison of recognized vs. misrecognized utterances indicates that F0 excursion, loudness and longer duration are significant prosodic characteristics of both WA- and CA-defined failed recognition attempts. Results from a set of machine learning experiments further show that prosodic differences can in fact be used to improve the prediction of misrecognitions with a high degree of accuracy for WA-based misrecognitions—

and an even higher degree when combined with information currently available from ASR systems. For example, the use of ASR confidence scores in TOOT predicted WA-defined misrecognitions with an error rate of 18.91%, while adding prosodic and other ASR features reduced the error to 8.64%. The additional power provided by prosodic information is more than simply a correlation of longer turns with higher recognition error, as our results show when we remove this information without significant degradation in predictive power; in particular, automatically calculated timing information performs much more poorly as a predictor for our W99 corpus than other prosodic features. For WA-defined misrecognitions, the differences between using simple acoustic-based confidence scores and our richer feature sets are quite considerable. For CA-defined misrecognitions in TOOT and WA-defined misrecognitions in W99, the improvement provided by prosodic features is significant but less dramatic. One of our future research directions will be to understand this difference.

Another future direction will be to address the issue of just **why** prosodic features provide such useful indicators of recognition failure. Does the occurrence of these features themselves make recognition difficult, or are they instead indirect correlates of other phenomena not captured in our study? While the negative influence of speaking rate variation on ASR has been reported before (e.g., Ostendorf et al., 1997), it is traditionally assumed that ASR is impervious to differences in F0 and RMS; yet, it is known that F0 and RMS variations co-vary to some extent with spectral characteristics (e.g., Swerts and Veldhuis, 1997; Fant et al., 1995), so that it is not unlikely that utterances with extreme values for these may differ critically from the training data. Other prosodic features may be more indirect indicators of errors. Longer utterances may simply provide more chance for error than shorter ones, while speakers who pause longer before utterances and take more time making them may also produce more disfluencies than others.

We have also extended our TOOT corpus analysis to include prosodic analyses of turns in which users become aware of misrecognitions

(Litman et al., 2001) and correct them (Hirschberg et al., 2001). In addition, we are exploring whether prosodic differences can help explain the “goat” phenomenon—the fact that some voices are recognized much more poorly than others (Doddington et al., 1998; Hirschberg et al., 1999). Our ultimate goal is to provide prosodically based mechanisms for identifying and reacting to ASR failures in spoken dialogue systems.

### Acknowledgements

We thank Mazim Rahim for providing access to and help in using the W99 corpus. Marc Swerts is partly sponsored by the Fund for Scientific Research–Flanders (FWO–Flanders) and the Netherlands Organization for Scientific Research (NWO).

### References

- Ammicht, E., Potamianos, A., Fosler-Lussier, E., 2001. Ambiguity representation and resolution in spoken dialogue systems. In: Proc. EUROSPEECH-01, Aalborg, pp. 2217–2220.
- Andorno, M., Laface, P., Gemello, R., 2002. Experiments in confidence scoring for word and sentence verification. In: Proc. Internat. Conf. on Spoken Language Processing-02, Denver, pp. 1377–1381.
- Bell, L., Gustafson, J., 1999. Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. In: Proc. Internat. Congress of Phonetic Sciences-99, San Francisco, pp. 1221–1224.
- Blaauw, E., 1992. Phonetic differences between read and spontaneous speech. In: Proc. Internat. Conf. on Spoken Language Processing-92, Banff, Vol. 1, pp. 751–758.
- Bouwman, A.G., Sturm, J., Boves, L., 1999. Incorporating confidence measures in the Dutch train timetable information system developed in the ARISE project. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Phoenix, Vol. 1, pp. 493–496.
- Bruce, G., 1995. Modelling Swedish intonation for read and spontaneous speech. In: Proc. Internat. Congress of Phonetic Sciences, Stockholm, Vol. 2, pp. 28–35.
- Cohen, W., 1996. Learning trees and rules with set-valued features. In: 14th Conference of the American Association of Artificial Intelligence, AAAI, Portland, pp. 709–716.
- Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D., 1998. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Proc. Internat. Conf. on Spoken Language Processing-98, Sydney, pp. 608–611.

- Falavigna, D., Gretter, R., Riccardi, G., 2002. Acoustic and word lattice based algorithms for confidence scores. In: Proc. Internat. Conf. on Spoken Language Processing-02, Denver, pp. 1621–1624.
- Fant, G., Liljencrants, J., Karlsson, I., Båvegård, M., 1995. Time and frequency domain aspects of voice source modelling. BR Speechmaps 6975, ESPRIT. Deliverable 27 WP 1.3.
- Guillevic, D., Gandrabur, S., Normandin, Y., 2002. Robust semantic confidence scoring. In: Proc. Internat. Conf. on Spoken Language Processing-02, Denver, pp. 853–856.
- Hirose, K., 1997. Disambiguating recognition results by prosodic features. In: Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer, pp. 327–342.
- Hirschberg, J., 1991. Using text analysis to predict international boundaries. In: Proc. Second European Conference on Speech Communication and Technology, Genova, pp. 1275–1278.
- Hirschberg, J., 1995. Prosodic and other acoustic cues to speaking style in spontaneous and read speech. In: Proc. Internat. Congress of Phonetic Sciences, Stockholm, Vol. 2, pp. 36–43.
- Hirschberg, J., Litman, D., Swerts, M., 1999. Prosodic cues to recognition errors. In: Proc. Automatic Speech Recognition and Understanding Workshop (ASRU'99), Keystone, pp. 349–352.
- Hirschberg, J., Litman, D., Swerts, M., 2001. Identifying user corrections automatically in spoken dialogue systems. In: Proc. NAACL-01, Pittsburgh, pp. 208–215.
- Kamm, C., Narayanan, S., Dutton, D., Ritenour, R., 1997. Evaluating spoken dialog systems for telecommunication services. In: Proc. EUROSPEECH-97, Rhodes, pp. 2203–2206.
- Kraayeveld, H., 1997. Idiosyncrasy in prosody. Speaker and speaker group identification in Dutch using melodic and temporal information. Ph.D. thesis, Nijmegen University.
- Krahmer, E., Swerts, M., Theune, M., Weegels, M., 2001. Error detection in spoken human–machine interaction. *International Journal of Speech Technology* 4 (1), 19–30.
- Levow, G.-A., 1998. Characterizing and recognizing spoken corrections in human–computer dialogue. In: Proc. 36th Annual Meeting of the Association of Computational Linguistics, COLING/ACL 98, Montreal, pp. 736–742.
- Litman, D., Pan, S., 1999. Empirically evaluating an adaptable spoken dialogue system. In: Proc. 7th International Conference on User Modeling (UM), Banff, pp. 55–64.
- Litman, D., Walker, M., Kearns, M., 1999. Automatic detection of poor speech recognition at the dialogue level. In: Proc. 37th Annual Meeting of the Association of Computational Linguistics, ACL99, College Park, pp. 309–316.
- Litman, D., Hirschberg, J., Swerts, M., 2001. Predicting user reactions to system error. In: Proc. ACL-2001, Toulouse, pp. 329–369.
- Moreno, P.J., Logan, B., Raj, B., 2001. A boosting approach for confidence scoring. In: Proc. EUROSPEECH-01, Aalborg, pp. 2109–2112.
- Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Guna-wardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfeld, T., 1997. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. Report on 1996 CLSP/JHU Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition.
- Oviatt, S.L., Levow, G., MacEachern, M., Kuhn, K., 1996. Modeling hyperarticulate speech during human–computer error resolution. In: Proc. Internat. Conf. on Spoken Language Processing-96, Philadelphia, pp. 801–804.
- Rahim, M., Pieraccini, R., Eckert, W., Levin, E., Di Fabbrizio, G., Riccardi, G., Lin, C., Kamm, C., 1999. W99—a spoken dialogue system for the ASRU'99 workshop. In: Proc. ASRU'99, Keystone.
- Sharp, R.D., Bocchieri, E., Castillo, C., Parthasarathy, S., Rath, C., Riley, M., Rowland, J., 1997. The Watson speech recognition engine. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing-97, Munich, pp. 4065–4068.
- Sohtau, H., Waibel, A., 1998. On the influence of hyperarticulated speech on recognition performance. In: Proc. Internat. Conf. on Spoken Language Processing-98, Sydney, pp. 225–228.
- Sohtau, H., Waibel, A., 2000. Specialized acoustic models for hyperarticulated speech. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing 2000, Istanbul, pp. 1779–1782.
- Sohtau, H., Metze, H., Waibel, A., 2002. Compensating for hyperarticulation by modeling articulatory properties. In: Proc. Internat. Conf. on Spoken Language Processing-02, Denver, pp. 83–86.
- Swerts, M., Ostendorf, M., 1997. Prosodic and lexical indications of discourse structure in human–machine interactions. *Speech Communication* 22, 25–41.
- Swerts, M., Veldhuis, R., 1997. Interactions between intonation and glottal-pulse characteristics. In: Botinis, A., Kouroupetoglou, G., Carayiannis, G., (Eds.), *Intonation: Theory, Models and Applications*, Athens, pp. 297–300.
- Swerts, M., Litman, D., Hirschberg, J., 2000. Corrections in spoken dialogue systems. In: Proc. Internat. Conf. on Spoken Language Processing-00, Beijing, pp. 615–618.
- Talkin, D., 1995. A Robust algorithm for pitch tracking (RAPT). In: Klein, W.B., Paliwal, K.K. (Eds.), *Speech Coding and Synthesis*, Athens, pp. 495–518.
- Veilleux, N., 1994. Computational models of the prosody/syntax mapping for spoken language Systems. Ph.D. thesis, Boston University.
- Wade, E., Shriberg, E.E., Price, P.J., 1992. User behaviors affecting speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing-92, Banff, Vol. 2, pp. 995–998.
- Walker, M., Fromer, J., Narayanan, S., 1998. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In: Proc. ACL/COLING, Montreal, pp. 1345–1352.
- Walker, M., Kamm, C., Litman, D., 2000a. Towards developing general models of usability with PARADISE. *Natural*

- Language Engineering: Special Issue on Best Practice in Spoken Language Dialogue System Engineering, Vol. 6, pp. 363–377.
- Walker, M., Langkilde, I., Wright, J., Gorin, A., Litman, D., 2000b. Learning to predict problematic situations in a spoken dialogue system: Experiments with How may I help you? In: Proc. NAACL-00, Seattle, pp. 210–217.
- Wang, H.-M., Lin, Y.-C., 2002. Error-tolerant spoken language understanding with confidence measuring. In: Proc. Internat. Conf. on Spoken Language Processing-02, Denver, pp. 1625–1628.
- Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. In: Proc. Internat. Conf. on Spoken Language Processing-96, Philadelphia, pp. S16–S19 (addendum).
- Zeljko, I., 1996. Decoding optimal state sequences with smooth state likelihoods. In: International Conference on Acoustics, Speech, and Signal Processing 96, Atlanta, pp. 129–132.
- Zhang, R., Rudnicky, A., 2001. Word level confidence annotation using combinations of features. In: Proc. EURO-SPEECH-01, Aalborg, pp. 2105–2108.