PROSODIC CUES TO RECOGNITION ERRORS

Julia Hirschberg Diane Litman

AT&T Labs — Research
Florham Park, NJ
{julia,diane}@research.att.com

Marc Swerts

IPO, Center for User-System Interaction
Eindhoven, The Netherlands
m.g.j.swerts@tue.nl

ABSTRACT

We identify methods of distinguishing between correctly and incorrectly recognized utterances (scored by hand for semantic concept accuracy) for a speech recognition system, using acoustic/prosodic characteristics. The analysis was performed on data collected during independent experiments done with an interactive voice response system that provides travel information over the phone.

1. INTRODUCTION

There has been little research in the field of automatic speech recognition (ASR) on the question of how misrecognized utterances differ from correctly recognized utterances. Recognition performance is known to vary depending upon the relative formality or casualness of speaking style [14], but there has been little attempt to identify this variation precisely. An exception is a study of the effect of speaking style on recognition performance in the Switchboard Corpus in which a standard recognition system was augmented with a conditioning variable, the speaking style (mode) [8]. Lexical and prosodic features hypothesized as possible correlates of acoustic modeling errors were calculated for correctly and incorrectly modeled regions. The best automatically trained procedure distinguished correct from incorrect matches 24.4% of the time (baseline: 35.8%), with the most important feature being speaking rate. Prosody has also been shown to be effective in ranking recognition hypotheses, as a post-processing filter which degrades wrong hypotheses and selects the correct ones [5, 12, 4]. Other research has identified particular types of speech variability, both across and within speakers, which may also be linked to prosodic variation, and which have been shown to have a negative impact on recognition performance. Several studies [13, 9, 11, 6, 1] report that hyperarticulate 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 shown that such speech also decreases recognition performance [10].

Our research investigates the characteristics of correctly recognized vs. misrecognized speaker turns in an Interactive Voice Response (IVR) system. We present results of prosodic analyses of these turns which show that in fact these turns differ significantly in terms of a number of prosodic features. We also describe machine learning experiments that combine prosodic features with additional sources of automatically obtainable information, such as recognized string and acoustic confidence score, to predict whether speaker turns will be correctly recognized or not with a high degree of accuracy.

2. CORPUS

Our corpus consists of a set of dialogues between humans and TOOT, a spoken dialogue system for accessing train schedules from the web via telephone, which was collected to study variations in IVR strategy [7]. TOOT is implemented using an IVR platform developed at AT&T, combining ASR, text-to-speech, a phone interface, and modules for specifying a finite-state machine dialogue manager and application functions. 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 at any dialogue state.

Subjects performed four tasks with one of several versions of TOOT, that differed in terms of locus of initiative (system, user, or mixed), confirmation strategy (explicit, implicit, or none), and whether these conditions could be changed by the user during the task. Subjects were 39 students, 20 native speakers and 19 non-native speakers; 16 subjects were female and 23 male. Dialogues were recorded and system and user behavior logged automatically. The *concept accuracy* (CA) of each (non-rejected) turn was manually labeled by one of the experimenters, by listening to the dialogue recordings while examining the system log. If the ASR output correctly captured all the task-related information in the turn (e.g. time, departure and arrival cities), the turn was given a CA score of 1 (a *semantically correct recognition*). Otherwise, the CA score reflected the percentage of correctly recognized task information in the turn. For the study described below, we examined 2067 user turns from 152 dialogues in this corpus, which included 170 rejections (based on acoustic confidence scores [15]) and 491 turns with CA < 1.

3. CHARACTERIZING MISRECOGNITIONS

Our goal was to determine whether misrecognized turns differed from recognized turns in terms of their prosodic features, and, if so, to develop methods to distinguish these automatically. To this end, we examined the f0 and rms average and maximum values for each turn (calculated from the output of Entropic Research Laboratory's pitch tracker, get_f0); the total duration of the user turn; the length of pause preceding the turn; the amount of silence within the turn (percentage of zero frames); and the tempo (calculated as syllables (from the recognized string) per second). For each feature, we calculated mean values for recognized (concept accuracy score of 1) turns and of misrecognized (concept accuracy less than 1) turns on a per speaker basis. We then performed paired t-tests on these means (Table 1), to ensure that our results applied across speaker. Our results indicate that misrecognized turns do differ in important ways from correctly recognized ones. We found statistically significant differences (at a 95% confidence level, p$\leq .05$) across speakers, for the first five features listed in the table. Misrecognized turns exhibited significantly higher f0 and rms maxima, lower mean rms, longer durations, and were preceded by longer pauses than were correctly recognized speaker turns. Speaking rate, f0 average, and amount of silence within turn showed no such differences. We ran similar tests on normalized versions of all features, normalizing by first turn in task or by preceding turn or by scaling values by speaker, with similar results in each case.

Table 1: Comparison of Misrecognized vs. Recognized Turns by Prosodic Feature Across Speakers

Feature	T-stat	Mean Misrecognized - Recognized	P
F0 Max	5.40	28.87 Hz	0
RMS Max	3.00	185.74	.005
RMS Avg	2.07	-32.43	.05
Duration	10.42	2.34 sec	0
Prior Pause	5.22	.38 sec	0
Tempo	.45	.13 sps	.65
F0 Avg	1.36	1.63 Hz	.18
% Silence	.05	-.02%	.29

When we include rejected turns with misrecognized turns, and simply compare these non-recognitions with correct recognitions, we also get statistically significant differences in the same direction for f0 and rms maximum, duration, and preceding pause, as above. So, our findings in fact distinguish all 'problem' turns, including those the ASR engine itself hypothesizes (rightly or wrongly) are poorly recognized.

The features we found to be significant indicators of a failed recognition, f0 excursion, loudness, long prior pause, and longer duration, are all features previously associated with hyperarticulated speech. Since prior research has suggested that speakers may respond to failed recognition attempts with hyperarticulation, which itself may lead to more recognition failures, had we in fact identified a means of characterizing and identifying hyperarticulated speech prosodically? Since we had independently labeled all speaker turns for evidence of hyperarticulation (two of the authors labeled each turn as 0, 'not hyperarticulated', 1 'some hyperarticulation in the turn', 2 'hyperarticulated', following [13]), we were able to test this possibility. We excluded any turn either labeler had labeled as '1' or '2', and again performed paired t-tests on mean values of misrecognized vs. recognized turns for each speaker. Results again showed that f0 and rms maximum, total turn duration, and prior pause are significantly different for misrecognized vs. recognized turns. So, while, indeed, hyperarticulated turns are recognized significantly more poorly than non-hyperarticulated, and misrecognized turns are significantly more likely to have a higher hyperarticulation score, and while hyperarticulated turns do have a significantly higher f0 maximum, rms maximum, and overall duration than non-hyperarticulated turns, our findings for the prosodic characteristics of misrecognized and of misrecognized turns hold even when perceptibly hyperarticulated turns are excluded from the corpus.

While the prosodic features we found characterizing semantic misrecognitions hold across speakers, we also checked to see whether in fact these features could help explain the 'goat' phenomenon in ASR experience — the fact that some voices are recognized much more poorly than others [3]. Could this phenomenon be accounted for in terms of a propensity for some speakers to typically employ greater f0 and rms excursion, to produce turns of a longer duration, and to produce prolonged pauses before their turns? To explore this possibility, we divided our speakers into those whose ratio of unrecognized to recognized turns was .5 or greater (goats) and those with a ratio of less than .5 (sheep). Of 18 goats overall, 7 were women, 9 were non-native speakers, and 14 of the total fell into at least one of those categories. We restricted our analysis to speakers participating in experimental conditions in which the system provided implicit confirmation and permitted mixed speaker/system initiative, since this condition provided roughly equal numbers of goats (8) and sheep (5). Three of the 8 goats were non-native speakers, and 2 of the sheep, so the proportion of non-native to native speakers was roughly the same in each category. Of the goats, 3 were male and 5 female, while only one of the 5 sheep was female.

We calculated means for each speaker over all turns for the features f0 and rms maximum, turn duration, and prior pause, which we found above to characterize unrecognized vs. recognized turns, and performed t-tests on the means. The results do indeed suggest that goats may be distinguished from sheep prosodically, although not in the same way that unrecognized turns can be distinguished. Our 'goats' are distinguished from better recognized speakers in terms of significant differences in overall f0 maximum and tempo – the latter a factor that does not figure in our prior analyses; goats exhibit larger f0 excursions and speak more slowly than sheep. However, when prosodic features are normalized by first turn in task or by mean for speaker over task, f0 maximum ceases to be a significant predictor of 'goat' status. So, apparently, speakers with higher ranges prove difficult for the recognizer, as previous folk wisdom has found. But the simple hypothesis that goats more often produce utterances with the prosodic characteristics we find distinguish misrecognize utterances is not supported.

4. MACHINE LEARNING EXPERIMENTS

Given the prosodic differences between misrecognized and correctly recognized utterances in our corpus, is it possible to predict accurately when a particular utterance will be semantically misrecognized or not? This section describes experiments using the machine learning program RIPPER [2] to automatically induce such a prediction model. Like many learning programs, RIPPER takes as input the classes to be learned, the names and possible values (symbolic, continuous, or text) of a set of features, 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 greedy search guided by an information gain metric, and is expressed as an ordered set of if-then rules. When multiple rules are applicable, RIPPER uses the first rule.

Our 2 predicted classes correspond to concept accuracy of 1 (T) or not (F). As in Section 3, the features used in our learning experiments include the raw acoustic features. However, our learning experiments also include other potential predictors of misrecognition. Four features represent the experimental conditions described in Section 2 (system (adaptive or non), dialogue strategy (mixed initiative with implicit system confirmation, system initiative with explicit confirmation, and user initiative with no system confirmation), subject, and task number), while 2 features represent the ASR output (confidence score and recognized string). All of these features could in principle be made automatically available to a spoken dialogue system. We also include one hand-labeled prosodic feature, hyperarticulation score, described in Section 3.

Table 2 shows the relative performance of some feature sets we tried.[1] The best performing feature set uses all of

Table 2: Estimated Error for Predicting Misrecognitions

Features Used	Error (SE)
All	10.79% (.83)
Prosody, ASR Conf	12.71% (.75)
ASR Conf, ASR String	13.39% (.85)
ASR String	14.41% (.78)
Prosody	16.15% (.87)
ASR Conf	18.30% (.81)
Baseline	25.88%

the features discussed above, and gives an error of 10.79% +/- .83%. Furthermore, even when we exclude the features that were not automatically logged in the original experiment (even if they could have been automated, in principle), or that represent the experimental conditions, our results are statistically equivalent. In contrast, a baseline classifier predicting the majority class of T has an error of 25.88%. A sample classification model (error of 11.21% +/- .72%) learned from only the prosodic and other automatically available features is shown below:

if (asr $\leq$ -3.33589) $\wedge$ (dur $\geq$ 1.42496) $\wedge$ (tempo $\leq$ 2.54345) **then** *F*
if (asr $\leq$ -3.80883) $\wedge$ (dur $\geq$ 1.13314) $\wedge$ (f0av $\leq$ 201.595) **then** *F*
if (asr $\leq$ -2.75271) $\wedge$ (tempo $\leq$ 1.51501) $\wedge$ (zeros $\leq$ 0.595294) **then** *F*
if (asr $\leq$ -2.74865) $\wedge$ (string contains 'help') **then** *F*
if (asr $\leq$ -5.04968) $\wedge$ (dur $\geq$ 0.925401) $\wedge$ (zeros $\leq$ 0.453237) **then** *F*
if (asr $\leq$ -2.71435) $\wedge$ (dur $\geq$ 0.982058) $\wedge$ (f0max $\leq$ 155.887) **then** *F*
if (asr $\leq$ -3.61652) $\wedge$ (string contains '8') **then** *F*
if (dur $\geq$ 1.39754) $\wedge$ (tempo $\leq$ 2.25751) $\wedge$ (zeros $\leq$ 0.626667) **then** *F*
if tempo $\geq$ 7.25392 **then** *F*
else T

Note that 5 prosodic features appear in these rules. Although f0 maximum and duration were shown to be significant features in our earlier t-tests, these rules also include 3 additional prosodic features (tempo, f0 average and % silence (zeros)). If we consider the predictive power of each of our prosodic features in isolation, we find that duration is the best performing feature (error of 17.01% +/- .99%), followed by tempo (20.70% +/- .99%). All of the other prosodic features perform no better than the baseline.

In addition to the prosodic features, this ruleset also contains two ASR features - confidence score (asr) and recognized string (string). While confidence score was intended to predict word accuracy, it also appears useful in predicting concept accuracy. More surprising perhaps is the appearance of certain particular recognized strings as cues to

[1]The error rates and standard errors (SE) result from 25-fold cross-validation. Two error rates are statistically significantly different when the errors plus or minus twice the standard error do not overlap.

recognition accuracy. Note the use of the recognized string "help" in the fourth rule and "8" in the seventh. Using ASR string alone in fact is the best single feature in predicting our data (14.41%), although this error rate is still worse than our best combination of features, and is probably less likely to generalize. Finally, despite the fact that system versions and dialogue strategies differ in our corpus, none of these features appear in the rules.

From these results it is still not clear what the contribution of our prosodic variables is. Do they simply correlate with information already available to recognition systems — in particular, are they adding any predictive power not already available from the ASR confidence scores? To investigate this possibility, we examined the contribution of the confidence score alone in predicting misrecognitions vs. correct recognitions: using this metric alone, the estimated error is 18.3%, significantly worse than the error estimated when our prosodic features are also included (12.71%), and statistically equivalent to using prosody alone (16.15%). But using confidence score plus recognized string gives an estimated error of 13.39%, which, at a 95% confidence level, is not significantly different from our best performing feature set. With respect to the prosodic features, it should be noted that when we compare the performance of prosodic features normalized in different ways (discussed in Section 3), we find that the unnormalized feature set significantly outperforms (by 6-10%) the prosodic features normalized by first task utterance or by previous utterance. This difference may indicate that there are limits on the ranges in features such as f0 and rms maximum, duration and preceding pause within which recognition performance is optimal.

5. DISCUSSION

Results from a statistical comparison of recognized vs. misrecognized utterances indicate that there are significant prosodic differences between misrecognized and correctly recognized utterances, both within and across speakers. Our machine learning experiments show that prosodic features can be used to predict semantic misrecognitions with a high degree of accuracy, improving over the use of simple acoustic-based confidence scores by about one third. Next steps include testing whether our results hold up across recognizers, continuing our comparison of goats vs. sheep, and extending our current study characterizing misrecognitions to include analysis of turns in which users become aware of these misrecognitions and turns in which they correct them. Our ultimate goal is to provide prosodically-based mechanisms for identifying and reacting to ASR failures in IVR systems.

6. REFERENCES

[1] L. Bell and J. Gustafson. Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. In *Proc. ICPhS-99*.

[2] W. Cohen. Learning trees and rules with set-valued features. In *Proc. AAAI-96*.

[3] G. Doddington et al. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proc. ICSLP-98*.

[4] K. Hirose. Disambiguating recognition results by prosodic features. In *Computing Prosody: Computational Models for Processing Spontaneous Speech*, pages 327–342. 1997.

[5] J. Hirschberg. Using text analysis to predict intonational boundaries. In *Proc. EUROSPEECH-91*.

[6] G. Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings COLING/ACL-98*.

[7] D. Litman and S. Pan. Empirically evaluating an adaptable spoken dialogue system. In *Proc. the 7th International Conference on User Modeling*, 1999.

[8] M. Ostendorf et al. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. 1996 CLSP/JHU Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition report, 1997.

[9] S. Oviatt et al. Modeling hyperarticulate speech during human-computer error resolution. In *Proc. ICSLP-96*.

[10] H. Soltau and A. Waibel. On the influence of hyperarticulated speech on recognition performance. In *Proc. ICSLP-98*.

[11] M. Swerts and M. Ostendorf. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22, 1997.

[12] N. Veilleux. *Computational Models of the Prosody/Syntax Mapping for Spoken Language Systems*. PhD thesis, Boston University, 1994.

[13] E. Wade et al. User behaviors affecting speech recognition. In *Proc. ICSLP-92*.

[14] M. Weintrab et al. Effect of speaking style on LVCSR performance. In *Proc. ICSLP-96*.

[15] I. Zeljkovic. Decoding optimal state sequences with smooth state likelihoods. In *Proc. ICASSP-96*.