6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

GENERALIZING PROSODIC PREDICTION OF SPEECH RECOGNITION ERRORS

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

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

ABSTRACT

Since users of spoken dialogue systems have difficulty correcting system misconceptions, it is important for automatic speech recognition (ASR) systems to know when their best hypothesis is incorrect. We compare results of previous experiments which showed that prosody improves the detection of ASR errors to experiments with a new system and new domain, the W99 conference registration system. Our new results again show that prosodic features can improve prediction of ASR misrecognitions over the use of other standard techniques for ASR rejection.

1. INTRODUCTION

Users find it difficult to correct system misunderstandings (e.g., “When do you want to go to Boston?” when the user has said “Baltimore”) [4]. On the other hand, users are frustrated by unnecessary confirmations (“Did you say Baltimore?”) and rejections (“I didn’t understand you, can you please repeat?”). So it is important for systems to know when their best hypothesis is incorrect, so that they can better confirm or reject the user’s input [11], or, when many errors have occurred, change their interaction strategy [7].

In previous research we investigated the importance of a variety of prosodic and other cues to the automatic detection of misrecognitions in spoken dialogue systems [2, 5]. The data examined was obtained from subjects performing specified train information gathering tasks with TOOT, an experimental phone-based spoken dialogue system [6]. TOOT was implemented on a platform developed at AT&T combining ASR, text-to-speech, a phone interface, a finite-state dialogue manager, and application functions [3]. The speech recognizer employed in this platform, BLASR, 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 [3, 8]. In these studies, we found not only that there were major differences in F0 excursion, loudness, prior pause, and overall duration for user turns that were misrecognized vs. those that were correctly recognized, but also that machine learning techniques using these and other automatically available features, such as acoustic confidence score, recognized string, and grammar, could be employed to predict misrecognitions with a high degree of accuracy. Prosodic features could predict misrecognitions with only 12.76% error — and were even more accurate (6.53% error) when combined with acoustic confidence scores, identify of recognized string, and grammar state — all information currently available from ASR systems. These error rates compare quite favorably with 22.23% error when misrecognition is predicted using only the standard feature used for ASR rejection, acoustic confidence score.

We proposed that our accurate prediction might be due to our ability to detect, through prosodic and other cues, utterances not well modeled by the speech recognizer’s training corpus— i.e., too loud, too long, and so on. To determine the generalizability of our results, we repeated our experiments on a new corpus, in a new domain, with a different ASR engine. In this paper we report results of experiments on the W99 corpus [9], collected from a spoken dialogue system used for conference registration, which employed the Watson ASR system for recognition [10]. We compare earlier results for the TOOT corpus with results for this new W99 corpus, both with respect to features that significantly differentiate misrecognized from recognized data and to the overall performance of our machine learning techniques. Our new results again show significant prosodic correlates of ASR errors, and that prosodic features can help to improve prediction of ASR misrecognitions.

2. THE W99 CORPUS

W99 is a spoken dialogue system used to support registration and information access for the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU’99) [9]. It 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. W99 used WATSON [10], a speaker-independent hidden Markov model ASR system, with HMMs trained using maximum likelihood estimation followed by minimum classification error training. W99 rejected utterances based on ASR confidence score. As with the TOOT platform, ASR confidence scores were available only at the turn, not the word, level. W99 used a mixed initiative dialogue strategy: the system generally gave the user the initiative (e.g., users responded fluently and naturally 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 . . . ”). The initial version of W99 used acoustic models from a pre-existing call-routing application. State-dependent bigram language models were also obtained from this application, as well as from interactions collected using a text-only version of W99. The data examined in this paper consists of approximately 3000 utterances, obtained during both an experimental study evaluating W99 and during a data collection phase where callers were testing and exploring the system capabilities. Approximately three quarters of the utterances were obtained from male callers. The recognition results used in this paper are on-line results obtained during actual system use.

Overall, 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, se-

Feature	T-stat	Mean Misrecognized - Recognized	P	T-stat	Mean Misrecognized - Recognized	P
*#F0 Max	7.83	30.31 Hz	0	7.07	23.81 Hz	0
*F0 Mean	3.66	4.12 Hz	0	.06	-.10 Hz	.95
*#RMS Max	5.70	235.93	0	4.90	335.48	0
RMS Mean	-.57	-8.50	.57	.34	6.70	.7352
*#Duration	10.30	2.20 sec	0	10.55	1.88	0
#Tempo	-.05	.15 sps	.13	7.23	.28 sps	0
*#% Silence	-5.15	-.06%	0	9.58	-.06%	0
*TOOT/#W99 data significant at a 95% confidence level (p$\leq .05$)						

Table 1: Prosodic Features of Misrecognized (WER>0) vs. Recognized Turns: TOOT vs. W99.

mantic analysis) are different, with W99 using newer and generally more robust technology (e.g., stochastic language models). The TOOT data was obtained from structured experiments, while the W99 data included both experimental and non-experimental data. Finally, W99 uses a primarily user-initiative dialogue strategy with limited backoff to system-initiative, while TOOT employs three types of initiative and three types of confirmation strategies.

3. CHARACTERISTICS OF MISRECOGNITIONS

As in our previous work, we examined the W99 corpus to see which, if any, prosodic features distinguished misrecognitions from correctly recognized utterances. For TOOT we had hand-labeled concept accuracy (CA) scores as well as WER; while CA measures semantic accuracy, while WER measures word accuracy. For the W99 data we lacked accurate CA scores, so we present results only for WER-defined misrecognitions and compare those results only to the WER-defined case for TOOT. Our unit of analysis in both cases is the speaker turn. We divide our corpus into turns with WER > 0 (misrecognitions) and those with WER$= 0$ (correct recognitions). In TOOT we lacked the original speech files as segmented by the speech recognizer, so were forced to manually segment user turns from a recording of both sides of the interaction. For the W99 corpus, speaker turns are automatically end-pointed, so the user turns correspond to exactly what the ASR engine used for recognition.

For each speaker turn we examined the same prosodic features we had examined for the TOOT corpus: maximum and mean fundamental frequency values (F0 Max, F0 Mean); maximum and mean energy values (RMS Max, RMS Mean); total duration; speaking rate (Tempo); and amount of silence within the turn (% Silence). As before, F0 and RMS values were calculated from the output of Entropic Research Laboratory's pitch tracker, *get_f0*. Speaking rate was approximated in terms of syllables in the recognized string per second. % Silence was defined as the percentage of zero frames in the turn, i.e., roughly the percentage of time within the turn that the speaker was silent. Since the W99 data did not contain explicit speaker identification for a given session, we collapsed all data from all sessions into a single, temporally-ordered 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 1 compares prosodic characteristics for misrecognitions vs. correct recognitions for the original TOOT data and the new W99 corpus. These results indicate that misrecognized turns for the W99 data resemble those we previously examined for TOOT, in that misrecognized turns contain more extreme pitch excursions (higher F0 maximum), louder portions of speech (higher rms maximum), are longer, and contain less internal silence — all in comparison to correctly recognized turns.

These results are based on means for raw scores for prosodic features in each turn. For our TOOT experiments, little difference was found between raw scores and scores normalized by value of the speaker's first or of preceding turn, or by converting all a speaker's turns to *z* scores. However, for the W99 corpus, this picture is somewhat different. While means calculated on the absolute values for duration, rms maximum, F0 maximum, tempo, and percentage of internal silence distinguish misrecognitions from recognitions, when these values are normalized by preceding turn, only duration, F0 maximum and tempo significantly distinguish the two groups of turns.[1] This may indicate that there are limits on the ranges of prosodic features within which recognition performance is optimal. Thus, absolute deviation from some particular range, rather than relative differences in prosodic values, seem to be associated with recognition failures in W99.

4. PREDICTING MISRECOGNITIONS USING MACHINE LEARNING

This section describes experiments using the machine learning program RIPPER [1] to automatically induce models (using prosodic as well as additional features) for predicting misrecognitions in W99. RIPPER takes as input the classes to be learned, a set of feature names and possible values, and training data specifying the class and feature values for each training example. It outputs a classification model for predicting the class of future examples, expressed as an ordered set of if-then rules.

[1] However, since our W99 data did not include speaker identification, the normalization of the first turn in a dialogue by preceding turn sometimes was based on data from another speaker. Also, we could not normalize by first turn or scaling by speaker.

Features Used	Error	SE
All ASR, Prompt	22.77%	.59
Prosody, All ASR, Prompt	23.66%	.80
Prosody, String	23.70%	.63
Confidence, String, LM	23.77%	.87
All features	23.91%	.85
Prosody, Confidence, LM	24.07%	.83
Prosody, Confidence, String, LM	24.19%	.94
Prosody, Confidence	24.35%	.87
Confidence, LM	25.68%	.78
Confidence	26.14%	.80
Prosody	26.17%	.73
% Silence	31.30%	.93
Tempo	31.58%	.92
String	32.94%	.91
Prosody Normalized	36.31%	.79
Majority Class Baseline	39.67%	

Table 2: Estimated Error for Predicting Misrecognized Turns.

As in Section 3, our predicted classes correspond to correct recognition (T) or not (F), and each speaker turn is represented as a set of features. The features include the raw prosodic features described in Section 3 (which we will refer to as the feature set "Prosody") and six additional potential predictors of misrecognition. Four features are from the ASR process: LM (the dialogue state specific language model used to recognize the turn); Confidence (the turn-level acoustic confidence score output by the recognizer); String (the recognized string); and Likelihood (the normalized likelihood score from the decoder). We included these features as a baseline against which to test new methods of predicting misrecognitions, although a typical ASR system uses only confidence score in its rejection calculations. In addition, the feature Prompt represents the W99 prompt that preceded the user's turn, while Gender was labeled during the corpus transcription process (all other features were obtained automatically).

Table 2 shows the relative performance of a number of the feature sets we examined, and compares these results with a baseline classifier that predicts that ASR is always wrong (the majority class, F).[2] Our first interesting result is that the best performing feature set ("All ASR, Prompt", error of 22.77%) significantly outperforms the "standard" use of ASR confidence scores to determine misrecognitions ("Confidence", error of 26.14%). The best performing feature set, in contrast, includes the system prompt that generated the user's utterance, as well as the features arising from ASR processing (recognized string, language model, normalized likelihood, and confidence score). Note that the performance of the best feature set is statistically equivalent to that of the next seven feature sets (i.e., through "Prosody, Confidence").

if (confidence $\geq$ 910) $\wedge$ (string contains 'yes') **then** *T*
if (confidence $\geq$ 860) $\wedge$ (string contains 'no') **then** *T*
if (confidence $\geq$ 890) $\wedge$ (string contains 'yes') **then** *T*
if (confidence $\geq$ 860) $\wedge$ (language model $=$ help) **then** *T*
if (confidence $\geq$ 880) $\wedge$ (string contains 'zero') **then** *T*
if (confidence $\geq$ 860) $\wedge$ (string contains 'goodbye') **then** *T*
if (confidence $\geq$ 860) $\wedge$ (string contains 'transportation') **then** *T*
if (confidence $\geq$ 860) $\wedge$ (string contains 'registration') $\wedge$ (prompt $=$ do-not-understand) **then** *T*
if (confidence $\geq$ 860) $\wedge$ (string contains 'three') **then** *T*
if (confidence $\geq$ 850) $\wedge$ (string contains 'registration') **then** *T*
if (confidence $\geq$ 880) $\wedge$ (string contains 'sure') **then** *T*
if (confidence $\leq$ 390) $\wedge$ (string contains 'no') **then** *T*
else F

Figure 1: Best Performing Ruleset for Predicting Correctly Recognized Turns.

Turning now to prosody, 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%). Thus, prosodic features perform comparably to the traditional ASR practice, and using both types of features seems to be better than using either in isolation (although these latter results are not quite significant at the 95% confidence level). Similarly, the error using ASR confidence scores and language model (25.68%) and the error using prosody (26.17%) are reduced when the two feature sets are combined (24.07%). Another interesting finding 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 error is statistically equivalent to the error of the best performing feature set (22.77%). In contrast, the use of ASR confidence score has an error of 26.14%, which is statistically worse than 22.77%. A caveat here is that the string feature, like the language model, is less likely to generalize from application to application, but even the ASR confidence score will not generalize from recognizer to recognizer. Nevertheless, our findings suggest new features to be considered as a means of improving rejection performance in stable systems. Finally, using multiple prosodic features ("Prosody", 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. While using prosodic features in conjunction with non-prosodic features (e.g., "Prosody , String", error of 23.7%) seems to outperform the use of prosodic features alone (error of 26.17%), none of these improvements are statistically significant. Also note that the raw prosodic features ("Prosody") significantly outperform the normalized version ("Prosody Normalized") by 10%.

The classification model learned from the best performing feature set in Table 2 is shown in Figure 1. Rules are presented in order of importance in classifying data. When multiple rules are applicable, RIPPER uses the first rule. When no rules are applicable, the default rule is applied (which in this example predicts that the utterance is a misrecognition (F)). The first rule says that if the turn is recognized by ASR as including the string "yes" with an acoustic

[2] The errors and standard errors (SE) result from 25-fold cross-validation. When two errors plus or minus twice the standard error do not overlap, they are statistically significantly different.

if (% silence $\geq$ .95) $\wedge$ (duration $\leq$ 11.82) **then** *T*

if (% silence $\geq$ .97) $\wedge$ (tempo $\leq$ .34) **then** *T*

if (% silence $\geq$ .90) $\wedge$ (duration $\leq$ 6.54) $\wedge$ (RMS max $\geq$ 797.31) **then** *T*

if (% tempo $\leq$ .33) $\wedge$ (duration $\leq$ 6.33) **then** *T*

if (% silence geq .95) $\wedge$ (duration $\leq$ 13.52) **then** *T*

if (duration $\leq$ .97) **then** *T*

if (% silence $\geq$ .85) $\wedge$ (duration $\leq$ 4.01) **then** *T*

if (% tempo $\leq$ .20) $\wedge$ (F0 mean $\leq$ 112.285) **then** *T*

if (% silence $\geq$ .56) $\wedge$ (duration $\leq$ 2.34) $\wedge$ (RMS max $\geq$ 1385.04) **then** *T*

if (% silence $\geq$ .91) $\wedge$ (duration $\leq$ 11.49) $\wedge$ (236.01 $\leq$ F0 max $\leq$ 352.36 **then** *T*

else F

Figure 2: Best Performing Prosodic Ruleset for Predicting Correctly Recognized Turns.

confidence score $\geq$ 910, then predict a correct recognition.[3] The traditional ASR confidence score feature appears in all rules, recognized string appears in all but one rule, and language model and prompt each appear in one rule.

The classification model learned from the feature set "Prosody" is shown in Figure 2. The rules contain all of the prosodic features considered in our experiment, except for RMS mean. % Silence and tempo, which were the best performing prosodic features in isolation, appear in 7 and 3 rules, respectively. Duration, which was no better than the baseline in isolation, appears in 8 rules. Duration thus appears to be a useful predictive feature in conjunction with other prosodic features. Note that all of the features shown to be significant in our statistical analysis (Section 3) occur in the rules, as does the feature F0 mean.

It is interesting to compare the W99 results with our previous results on TOOT, which used older ASR technology and poorer performing acoustic and language models. Several results generalize across the TOOT and W99 experiments: the use of prosody and ASR confidence score is better for predicting misrecognitions than using confidence score alone, and the use of multiple prosodic features outperforms any single prosodic feature. In TOOT, however, the best single prosodic feature for predicting misrecognitions was duration, rather than % Silence as in W99. Also, in TOOT, using *only* prosody to predict misrecognitions significantly outperformed using ASR confidence score (either with or without language model) to predict misrecognitions, and the best performing feature set overall for predicting misrecognitions included prosody. We hypothesize that the utility of the prosodic features compared to traditional ASR rejection methods (26.17% vs. 26.14% error for W99, and 12.76% vs. 22.23% error for TOOT) is inversely related to the quality of the ASR models.

5. DISCUSSION

We have generalized results from previous experiments using prosody to improve ASR error prediction to a new domain and recognition system. For both TOOT and W99, misrecognitions are significantly higher in pitch, louder, longer, and have less internal silence than correctly recognized utterances. For both TOOT and W99, adding prosodic information improves (in absolute terms) prediction based on ASR confidence scores alone. Also, adding other ASR-obtained features improves prediction over confidence scores alone. Since our improvements come for ASR systems not well adapted to the task at hand, our results suggest that our methods may provide a useful alternative for both assessing and improving system performance.

6. ACKNOWLEDGEMENTS

Thanks to Mazin Rahim for the W99 corpus and help in using it.

7. REFERENCES

1. W. Cohen. Learning trees and rules with set-valued features. In *14th Conference of the American Association of Artificial Intelligence, AAAI*, 1996.
2. J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU'99)*, 1999.
3. C. Kamm, S. Narayanan, D. Dutton, and R. Ritenour. Evaluating spoken dialog systems for telecommunication services. In *5th European Conference on Speech Technology and Communication, EUROSPEECH-97*, Rhodes, 1997.
4. E. Krahmer, M. Swerts, M. Theune, and M. Weegels. Error spotting in human-machine interactions. In *Proceedings of EUROSPEECH-99*, 1999.
5. D. J. Litman, J. B. Hirschberg, and M. Swerts. Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the First Annual Meeting*, Seattle, May 2000. North American Association for Computational Linguistics (NAACL-00).
6. D. J. Litman and S. Pan. Empirically evaluating an adaptable spoken dialogue system. In *Proceedings of the 7th International Conference on User Modeling (UM)*, 1999.
7. D. J. Litman and S. Pan. Predicting and adapting to poor speech recognition in a spoken dialogue system. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence, AAAI-2000*, 2000.
8. L. Rabiner, B. Juang, and C. Lee. An overview of automatic speech recognition. In C. Lee, F. Soong, and K. Paliwal, editors, *Automatic Speech and Speaker Recognition, Advanced Topics*, pages 1–30. Kluwer Academic Publishers, 1996.
9. M. Rahim, R. Pieracini, W. Eckert, E. Levin, G. D. Fabbrizio, G. Riccardi, C. Lin, and C. Kamm. W99 - a spoken dialogue system for the asru'99 workshop. In *Proc. ASRU'99*, 1999.
10. R. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland. The watson speech recognition engine. In *Proc. ICASSP97*, pages 4065–4068, 1997.
11. R. W. Smith. An evaluation of strategies for selectively verifying utterance meanings in spoken natural language dialog. *International Journal of Human-Computer Studies*, 48:627–647, 1998.

[3] The confidence scores observed in our data ranged from 990 to 280.